THE GENOMIC BASIS OF HETEROSIS IN HIGH-YIELDING TRIPLOID HYBRIDS OF WILLOW (*SALIX* SPP.) BIOENERGY CROPS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Craig Henry Carlson

December 2018

THE GENOMIC BASIS OF HETEROSIS IN HIGH-YIELDING TRIPLOID HYBRIDS OF WILLOW (*SALIX* SPP.) BIOENERGY CROPS

Craig Henry Carlson, Ph. D.

Cornell University 2018

Sustainable, renewable energy production from lignocellulosic bioenergy crops such as shrub willow (*Salix* spp.), can help offset fossil fuel usage without competing with food crops. Hybridization is a key component in shrub willow breeding because hybrids often exhibit heterosis for yield. The genomic basis of heterosis in willow has not been well characterized, especially in high-yielding triploid hybrids. The primary objectives of this study were 1) to evaluate diverse diploid, triploid, and tetraploid families for a suite of traits important for biomass production, 2) to identify genomic regions linked to these traits, 3) to investigate the patterns gene expression in diploid and triploid hybrids, and 4) to correlate heterosis with gene expression in triploid hybrids. This work supports evidence that triploid hybrids of shrub willow exhibit heterosis for biomass growth traits. A high proportion of expression-level dominance and additive inheritance was observed among triploids, as well as *cis*- and *trans*-regulatory divergence. Importantly, nonadditive gene expression correlates with heterosis for biomass yield and growth traits among triploid hybrids. The extensive phenotypes collected will provide a database for future trait mapping and serve as a foundation for genomic selection. These results offer a unique perspective on the genomic basis of heterosis in high-yielding triploid hybrids, and will contribute to the growing genomic toolkit for the improvement of shrub willow as a bioenergy crop.

BIOGRAPHICAL SKETCH

Craig has had a life-long interest in plants, dating back to his first job at the age of 13 planting trees and shrubs for his hometown in rural North Dakota. After touring around the US in a rock band for a few years after graduating high school, he enrolled in the Department of Plant Sciences at North Dakota State University, where he earned his Bachelor's and Master's degrees in Plant Science. Throughout his undergraduate studies, Craig completed summer research internships with the Montana Department of Agriculture, The Arnold Arboretum of Harvard University, and The United States National Arboretum. He was also awarded a John Deere travel grant to study in India for a semester. In 2013, Craig joined Dr. Larry Smart's Lab in the Department of Plant Breeding and Genetics at Cornell University. Beyond his primary research, he has participated on a number of collaborative projects, such as genetic mapping in willow breeding populations, sex determination in the Salicaceae, population genomics of downy mildew in cucurbits, and leaf rust in willows. During his time at Cornell, Craig received the *Munger/Murphy Award in Plant Breeding* as well as the *Barbara McClintock Graduate Student Award* for his outstanding performance in the areas of scholarship, research, teaching, and outreach. Craig hopes to soon lead a woody plant improvement program.

This dissertation is dedicated to my wife, Lauren Carlson.

Without you, I'd be lost.

To You, Lo.

ACKNOWLEDGMENTS

Larry Smart was the best advisor I could ask for. He pushed me. I have failed many times over, but he always walked me through my failures, and pushed me to improve. I respect his loyalty to his family, students, employees, and the community. Not only have I become a more thoughtful scientist under his wing, I'm more confident as a scientist. Here's to Stone IPA at Zach's Bar and figuring things out. I would also like to acknowledge my Graduate Special Committee, Dr. Walter DeJong, Dr. Jeffry Doyle, and Dr. Michael Gore for their advice and support. I gratefully acknowledge our technical staff, Lauren Carlson, Dawn Fishback, Jane Petzoldt, and Rebecca Wilk for their year-round dedication in the field and lab, as well as excellent technical support from Curt Carter, Matthew Christiansen, Steve Gordner, and Mark Scott. I would like to acknowledge Yongwook Choi, Agnes Chan, and Chris Town at the J. Craig Venter Institute for their collaboration and support. This research was supported by U.S. Department of Energy Office of Science, Office of Biological and Environmental Research, grant DE-SC0008375, by the U.S. Department of Agriculture National Institute of Food and Agriculture, Agriculture and Food Research Initiative grant 2012-68005-19703 for the NEWBio Coordinated Agricultural Project, and by National Science Foundation Dimensions in Biodiversity grant DEB-1542486. Finally, I acknowledge my family, especially my dad, Bill Carlson, as well as Neal Holland and Adam Volz for the support through the years and nurturing my early interest in tree improvement.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

HETEROSIS IN CROP PLANTS – A REVIEW

1.1 Abstract

Heterosis refers to the instance when the progeny of a cross exhibits greater biomass, vigor, or yield than the mean performance or the better of the parents. Numerous genetic models have been proposed to explain this non-additive phenomenon, including dominance, overdominance, pseudo-overdominance, and epistasis. However, it is likely that most examples of heterosis is far more complex on a molecular level, involving a network of interactions. These models have been rigorously tested in model crops, but will not necessarily apply to highly heterozygous, outcrossing crops, like willow (*Salix* spp.). My interest is to understand the genomic basis of heterosis through transcriptomics. Like any trait measured in field trials, gene expression can also be nonadditive. It is at this level the roles of regulatory networks on heterosis can be assessed. Here, a review of heterosis in model and non-model species is presented, with particular emphasis on the inheritance and regulatory divergence patterns of gene expression.

1.2 Introduction

1.2.1 Hybridization

The hybridization of diverse genotypes within a species often results in increased vigor, total biomass, and/or yield in F_1 progeny, relative to the parents. This phenomenon was termed "heterosis" by George Shull in 1914 (Shull 1946) and further characterized by East (East 1936). Early observations of heterosis were described for the progeny of crosses of inbred lines of maize, which are highly homozygous, and produce F_1 progeny with much greater heterozygosity. Since the Green Revolution, researchers have exploited heterosis in a number of domesticated

crops, which has had major impact on yield improvements (Duvick 1999). Plant breeders have traditionally assumed an additive model of inheritance, such that complex features that vary between parents are expected to display phenotypes in the progeny that are intermediate of those of the parents. Most beneficial agronomic traits (e.g., biomass yield) are complex and can be incrementally improved by many cycles of careful hybridization, evaluation, and selection. For instance, grain yield in wheat is considered to be an additive trait, because there are often thousands of loci, each with small effects that have some impact yield. Stacking many positive-effect loci can be accomplished by crossing parents of contrasting genetic backgrounds and subsequently performing numerous rounds of selection and inbreeding in order to fix positive, large-effect alleles in a breeding population. However, mildly deleterious loci are not easily purged by inbreeding alone. The degree of heterosis observed is often related to the genetic differences between the parents, such that maize inbreds can be organized into heterotic groups, based on their pedigree.

1.2.2 Measures of Relatedness

While knowledge of plant pedigree is sufficient to make reasonable improvements, a deeper understanding of relatedness or kinship is required to maximize gain from selection. A simple metric of relatedness can be obtained by comparing shared features in a population of individuals. In a classical sense, these features are phenotypic traits which tend to be most informative of individual relatedness in an evolutionary context, like the size and shape of inflorescences or topology of pollen grains. Taxonomic characterization based on phenotype alone is typically only informative at the level of species, and a far less powerful tool to detect differences within and among allopatric or sympatric populations. Hybrid zones often complicate measures of relatedness based on phenotype alone (Hardig et al., 2000) and have likely

contributed many misidentifications of type specimens. Thus, if pedigree is unknown, additional tools are required to resolve any differences.

Today, relatedness can be obtained by interrogating variation in DNA sequence rather than phenotype. Once a laborious task to produce, modern sequence-based marker technologies have revolutionized nearly all fields of biology by improving the efficiency of marker generation and coverage at increasing affordability. Early examples of molecular markers were PCR-based and designed to exploit the numbers of conserved repeat sequences (< 6 bp) scattered throughout the genome, known as microsatellites or simple sequence repeats (SSRs). SSRs are multiallelic and have the advantage of being codominant. While SSRs are often highly polymorphic and can often distinguish closely-related individuals, they are expensive, because they have low throughput and require previous sequence knowledge. These markers are still useful, but they generally lack the high-throughput and high-coverage requirements for trait mapping desired by crop researchers today. Recent advancements in low-coverage sequencing of multi-sample barcoded DNA cut with restriction enzymes (e.g., *ApeK*I and *EcoT22*I) have allowed researchers to utilize single-nucleotide polymorphisms (SNPs) to estimate relative genetic distances. Now numbering in the many thousands to millions, SNP-based markers have resolved a number of taxonomic issues and provided better estimates of relatedness, because less bias is associated with markers evenly dispersed throughout the genome.

Knowledge of genome-wide DNA sequence can also inform researchers about the genomic relatedness of individuals. Stretches of DNA common to related individuals are called haplotypes and represent identity by descent. Long haplotype blocks in common among genotypes are indicative of a lack of recent recombination, whereas short haplotype blocks are indicative of more recent recombination. Inbreeding fixes these haplotype blocks, such that

selfing species tend to have longer haplotype blocks and outcrossing species tend to have shorter haplotype blocks. At any particular region along the genome, DNA sequences are more likely to be inherited together if those sequences are physically close to one another. For instance, two SNPs are thought to be *linked* if their recombination frequencies are indicative of genomic proximity. In related populations, entire genetic maps of physical chromosomes can be constructed by binning and ordering segregating SNP markers into linkage groups. No reference genome assembly is perfect, and genetics does not lie, so major improvements in reference assemblies can be made by incorporating genetic maps to inform scaffold order and orientation. The only caveat is that the population used to construct the genetic map must be closely related to the reference genotype.

1.2.3 Inbreeding Depression

The most common explanation of heterosis is related to inbreeding depression, which is caused by increased homozygosity in an individual. The probability that two alleles in an individual are identical by descent (IBD) (inbreeding coefficient) is important in understanding the nature of inbreeding depression. When an individual has a high inbreeding coefficient, lower vigor and fertility rates are generally more likely to be expressed in that individual, compared to those with a low inbreeding coefficient. Darwin's early experiments on hybridization showed evidence for the hypothesis that inbreeding is disadvantageous for the hybrid, because it lowered vigor and fertility (Darwin, 1876). Homozygosity of detrimental alleles resulting from inbreeding results in a negative phenotype, compared with the heterozygote. Deleterious mutant alleles are maintained at low levels in populations and subject to natural selection. If these mutations cause inbreeding depression or are lethal, then inbreeding and selection should purge these mutations from a population. Loci with a heterozygote advantage are maintained at higher,

intermediate frequencies, and are subject to balancing selection. However, in highly inbred populations, loci with a heterozygote advantage are not maintained because selection favors the better homozygote. Moreover, if heterosis is due to detrimental mutations, then purging these alleles would produce high-yielding individuals by reducing the mutational load, but if heterosis is due to high numbers of loci with heterozygote advantage, then breeding uniform, highly heterozygous individuals would be more desirable (Charlesworth and Willis, 2009).

1.3 Classic Genetic Theories

1.3.1 Dominance and Overdominance in Inbred Crops

Long-standing theories to explain the genetic basis for heterosis include the *dominance* or *complementation* theory, in which dominant factors from one parent complement deleterious recessive factors from the other parent, such that the progeny have a greater overall proportion of growth-promoting, dominant alleles. This theory is supported by the characterization of variation in genomic structure and local synteny among heterotic groups of maize (Fu and Dooner, 2002). Much of our knowledge of heterosis comes from maize, because inbred parents have been artificially selected for maximum combining ability, which enables more accurate quantitative phenotyping than in outcrossing species (Lippman and Zamir, 2007).

Heterosis is also theorized to arise from superior properties of heterozygosity at individual loci, what known as *overdominance*. This synergistic effect (allelic interactiosn) is supported by studies in Arabidopsis (Redei 1962), tomato (Vrebalov et al., 2002; Lippman and Zamir, 2007), and maize (Larièpe et al., 2012). In Arabidopsis, plants heterozygous for the *erecta* mutation displayed overdominance, which was found to be caused by a mutation in a receptor-like kinase that is involved in numerous aspects of plant growth. The most well-known instance of overdominance in tomato was heterosis for yield in the interspecific cross *Solanum*

lycopersicum × *S. pennellii*, where a single overdominant locus increased yield as much as 50% over the parents (Zamir 2001). Another instance of overdominance in tomato was for a fruit-ripening mutation in a MADS-box transcription factor, which prevents fruit from ripening, but heterozygotes ripened over time and remained firm. Although this case is not necessarily associated with increased biomass, heterosis is a quantitative trait defined by its agronomic benefits.

Pseudo-overdominance combines concepts from both the dominance and overdominance models. It describes the situation when two recessive mutations are linked, but in repulsion, and resembles overdominance because of the tight linkage between loci. While repulsion phase linkage shows partial or complete dominance in the pseudo-overdominance model, linkage is not a prerequisite in the overdominance model, because a single locus can achieve heterosis. Overdominance observed at *Tms5* and *Ghd8* loci in hybrid rice was attributed to pseudo-overdominance of genes in repulsion on the same chromosome (Huang et al., 2016).

Finally, there is also evidence that epistasis can explain heterosis in rice (Hua et al., 2003), which was attributed to mostly negative interactions between QTL. While there are cases in which a single theory can be applied and even a single locus can explain an example of yield heterosis, such as the *Single Flower Truss* gene in tomato (Krieger et al., 2010), many other studies highlight the complexity of the multigenic basis for heterosis within even a single crop species (Shi et al., 2011), which only scratches the surface in describing heterosis across the diversity of the plant kingdom.

1.3.2 Applying Models of Heterosis to Heterozygous Outcrossing Species and Polyploids

The classical dominance model for the genetic basis for heterosis drawn from breeding of diploid inbred crops clearly does not fully explain heterosis in heterozygous, outcrossing and

often polyploid species that is characteristic of perennial bioenergy crops, such as poplar, willow, or switchgrass. Natural populations of these species are highly heterozygous, which tends to mask deleterious recessive alleles in parental genotypes (Casler et al., 2007; Kelleher et al., 2007; Lin et al., 2009). In addition, there is often a high degree of inbreeding depression and/or genetic load in these crops. The greatest levels of heterosis often occur as a result of the hybridization of genetically diverse species (Foster and Shaw 1988; Bradshaw and Stettler 1995; Stettler et al., 1996; Kopp et al., 2002; Marron and Ceulemans 2006). Improved performance or heterosis in species hybrids can be explained through complementation of non-functional paralogs in one species by a functional or neo-functional paralog in the other species (Rodgers-Melnick et al., 2012) or via overdominance due to an accumulation of heterozygous loci (Birchler et al., 2005), which in *Populus* was correlated with hybrid vigor (Li and Wu 1996). Progressive heterosis is the instance where maximizing diverse genomes in a polyploid results in increasingly greater magnitudes of heterosis (Birchler 2010). Hybrids between tetraploid lines of alfalfa (*Medicago sativa*), potato (*Solanum tuberosum*), and maize display greater heterosis than their diploid equivalents and produce progressive heterosis in double-cross hybrids (Mok and Peloquin 1975; Groose et al., 1989; Bingham et al., 1994; Birchler et al., 2003; Riddle and Birchler 2008). One of the earliest examples of progressive heterosis is from East (1936), where he noted crosses between allotetraploid tobacco species, *Nicotiana tabacum* and *N. rustica*, "…showed more heterosis than any other crosses I have observed."

If substantial parental divergence is a prerequisite for heterosis, it may be assumed that their dosage effect in hybrid progeny should be equal to that of diploid or tetraploid parents with equal genomic contributions, thus assuming complementation of alleles is not biased when observed in equal proportions. Gene dosage must be considered in models of heterosis to explain

the many examples of hybrid vigor in triploids produced through the cross of a diploid with a tetraploid or through production of a diploid gamete. In maize, a strong response to ploidy was identified, where artificially produced triploids were superior to monoploids, diploids, and tetraploid derivatives for 13 traits (Riddle et al. 2006), which showed there is genetic variation for the response to ploidy changes. Triploid *Miscanthus* × *giganteus* produces greater yield than diploid *M. sinensis* (Clifton-Brown and Lewandowski 2002; Glowacka et al., 2010), and high-yielding triploid cultivars of sugar beet and cassava have been developed that outperform diploids (Savitsky 1962; Hornsey 1975; Sreekumari et al., 1999). In banana (*Musa* spp.), cultivars are the triploid progeny of *M. acuminata* × *M. balbisiana*, and the dosage of one or the other genome distinguishes dessert bananas from plantains, whereby yield is contributed by *M. acuminata* and stress tolerance by *M. balbisiana* (Robinson and Sauco 2010). Triploids have been sought in willow breeding, since they often display improved vigor (Zsuffa et al., 1984), and there are rare examples of fast-growing triploid *Populus* that arose through abnormal gamete formation (Bergstrom 1940; Einspahr et al., 1963; Benson and Einspahr 1967; Einspahr et al., 1968; Einspahr et al., 1970; Winton and Einspahr 1970; Bradshaw and Stettler 1993), including the ‘gigas’ triploid aspen identified in Sweden by one of the fathers of plant breeding, Herman Nilsson-Ehle (Müntzing 1936; Nilsson-Ehle 1936).

1.4 Molecular Mechanisms Underlying Heterosis

1.4.1 Gene Expression

A number of recent studies have focused on quantifying patterns of gene expression with heterosis for yield in intra- and inter-specific hybrids and polyploids. Gene expression is a quantitative trait that can be assessed like any other trait, and it is hypothesized that gene expression levels will fall into additive or nonadditive inheritance categories (Springer and

Stupar, 2007; He et al., 2010). These categories are: dominance (at the level of one of the parents), overdominance or underdominance (above or below the high or low parent, respectively), additive (between the high and low parent), or conserved (no difference between the parents or progeny) (Figure 1.1).

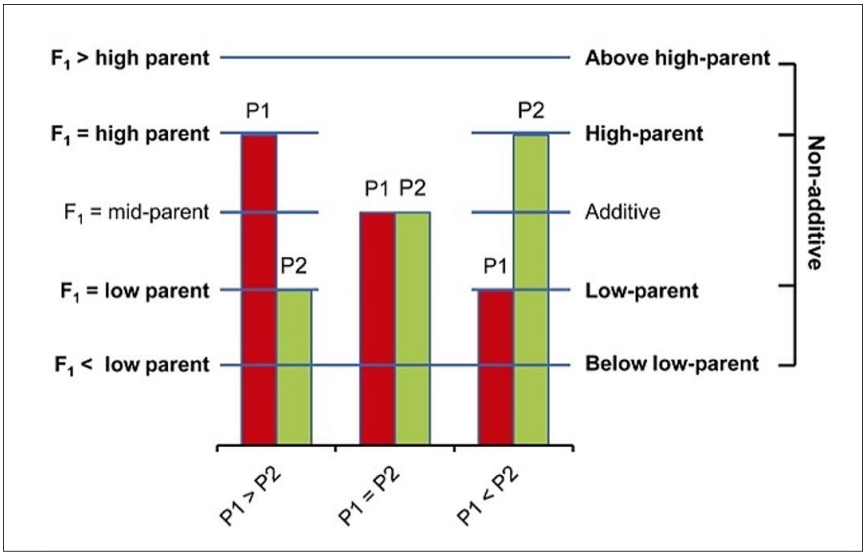

Figure 1.1 Diagram of the expected patterns of gene expression inheritance in species hybrids, taken from He et al. (2010).

The most common statistical method to derive these classifications are through Negative Binomial Exact tests for differential expression between the parents, as well as between the hybrid and both male and female parents. False Discovery Rates (FDR) control for false positives and are typically applied along with a minimum $\log_2$ fold-change ≥ 1 as a prerequisite. While the numbers of differentially expressed genes are attributed to more conserved expression using higher thresholds, the proportions do not change the biological significance of the results (Meiklejohn et al., 2014; Carlson et al., 2017). However, conservative thresholds reduce the likelihood of obtaining meaningful functional enrichments.

From early studies based on only a few dozen genes to recent research employing Illumina RNA-Seq to observe transcriptome-wide differences, a common result in maize, wheat,

and rice is that there is a high frequency of additive gene expression in hybrids (Sun et al., 2004; Guo et al., 2006; Stupar and Springer 2006; Swanson-Wagner et al., 2006; Stupar et al., 2008; Wei et al., 2009), although genes with nonadditive expression tended to display allele-specific expression (Guo et al. 2004; Springer and Stupar 2007a; Wei et al., 2009). This differential expression could be due to remote factors, whereby a small number of key regulatory genes can have a significant effect on heterosis (Ni et al., 2009; He et al., 2010; Goff 2011). It is believed that additive expression is a result of dosage-sensitivity. In maize, a greater proportion of nonadditive gene expression was observed in triploid and tetraploid hybrids with an effect of genome dosage (Guo et al., 1996; Auger et al., 2005; Birchler et al., 2005; Riddle et al., 2010). Gene expression, like any morphological trait, is affected by numerous loci, and crosses between parents that differ for traits usually exhibit intermediate or additive response (Tanksley 1993), which is analogous to dosage sensitivity.

1.4.2 Allele-Specific Expression

Recently, high-throughput RNA sequencing has been used to study the allelic imbalance of gene expression by comparing the expression of two alleles at a single locus. In contrast to microarray experiments, which is a subset of the transcriptome, RNA-Seq produces both isoform and expression data that is novel to each sample, which resolves issues of determining fold-differences in polymorphic sites and transcript abundance. Although RNA-Seq is not as stable as static DNA sequence data, RNA is closer to the product, based on the Central Dogma, DNA → RNA → Protein, and it is possible a transcript may contain more than one genetic variant on the same transcript. By using this approach, both transcript abundance and allelic bias can be used as a tool to elucidate biological questions with greater finesse than previous methods.

Gene expression is governed at the level of transcription initiation by interactions between *cis*- and *trans*-acting regulatory elements. Genes with strict *cis*-regulation, e.g., promoter sequences, have the same bias of expression of two alleles in both the hybrid and the parents and genes with strict *trans*-regulation, e.g., transcription factors and other regulatory elements, display allelic bias in the parents, but are expected to have equal levels of allelic expression in the hybrid, irrespective of the direction of fold-change. Allelic bias is a hallmark of the dominance hypothesis. Evidence from species hybrids of *Populus* found that a high proportion of the 30 genes assayed display allelic bias for the expression of genes from the two parents with evidence of novel alternative splicing patterns (Zhuang and Adams 2007; Scascitelli et al., 2010). Largely considered *cis*-acting, allelic bias may be confounded by nonadditive expression of a particular transcript due to *trans*-acting modifications. These differences are easily observable at the transcriptome-level, yet the occurrence of *cis* × *trans* effects (counteracting effects of *cis* and *trans* regulators of the same allele) is likewise very difficult to analyze *in silico*, because sound inference is dependent on sample-size, the method of assembly, variant calling, and the particular population under investigation. Furthermore, establishing a link between phenotypic heterosis and the genome and epigenome received from the parents is further confounded by a complex network of downstream metabolic responses driven by heterosis itself (Goff 2011).

A general trend is that *cis*-regulatory changes account for more of the divergent expression between more genetically divergent parents, whereas, significant *trans*-effects are thought to account for a higher proportion of the variation in gene expression between less-diverged parents (Wittkopp et al., 2008; Wittkopp and Kalay 2012). For instance, pure *cis*-effects imply the preservation of parental regulatory function and differential expression between

parents and hybrid are due to *trans*-effects caused by hybridization that brings two genomes together, which allows both alleles to be exposed to a common set of *trans*-acting elements (Figure 1.2).

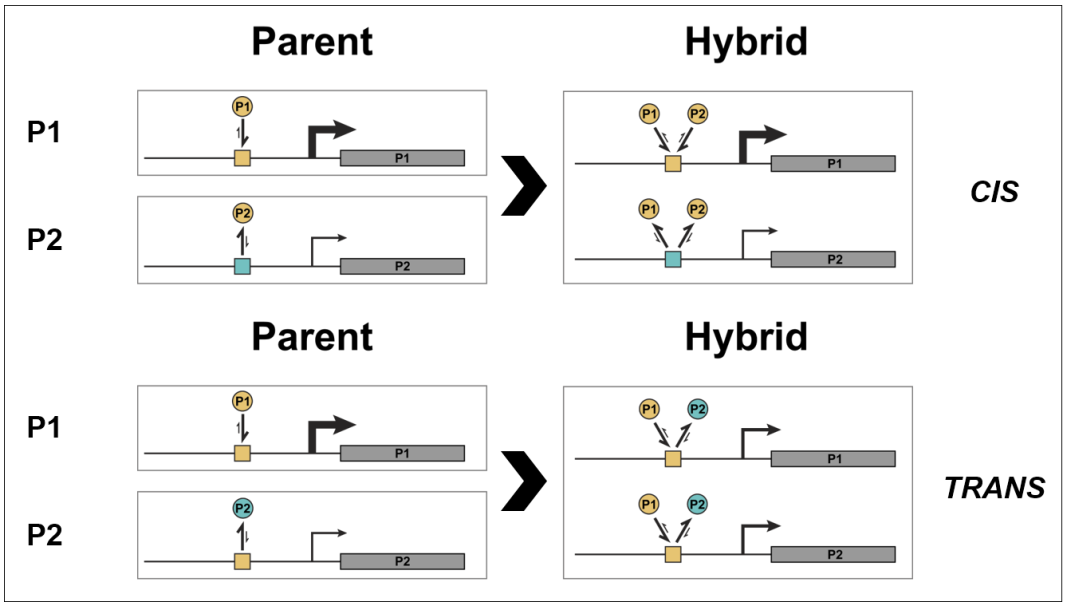

Figure 1.2 Diagram of *cis*- and *trans*-regulation. In the top panel, parent P1 has a common *trans*-factor (yellow circle) that has affinity to a *cis*-element (yellow square), but P2 does not share this affinity, due to polymorphism at its *cis*-element (cyan square), which results in differential expression of P1 and P2 genes in the parents and in the hybrid (*cis*-regulation). The bottom panel shows differential expression in the parents due to differing *trans*-factors, for a common cis-element, but the hybrid does not show differential expression of P1 and P2 copies, because P1 and P2 *trans*-factors are in a common background. Genes are represented as gray blocks and their expression levels as thickness of arrows. Figure adapted from Landry et al. (2005).

If this assumption is true, than in the interspecific hybrid, allelic imbalance (unequal expression of parental alleles) is a signature of *cis*-regulatory divergence, because both parental alleles in the hybrid are exposed to a common set of *trans*-acting regulators. Exceptions to this have been noted in cases of suspected population bottlenecks in one or both parental populations, where a higher than expected contribution of *trans*-variation might be the result of drift, rather than selection.

Cis-regulatory mutations are thought to account for instances of evolutionarily significant phenotypic change, but *trans*-regulatory evolution can also affect adaptive morphological change (Wray, 2007). Tests comparing parental and hybrid allele-specific expression (ASE) have been devised to discern between the *cis*- and *trans*-effects on nearby gene expression divergence. Wittkopp et al. (2004) showed functional *cis*-regulatory differences between closely-related *Drosophila* species by comparing the relative abundance of species-specific transcripts in F_1 hybrids and differences in *trans*-regulatory activity, inferred by comparing the ratio of allelic expression in hybrids with the ratio of allelic expression between species. Allele-specific expression tests in hybrids of inbred maize lines found that *cis*-acting regulatory variation accounted for the majority of the observed parental expression divergence and that *cis*-variation correlated with additive expression patterns in the F_1 hybrid (Stupar and Springer 2006). The level of domestication or selection biases within a particular pedigree and the contrasting patterns of linkage disequilibrium within populations play large roles in plant breeding systems, and are likely to impact gene expression in the hybrid by differences in specificity and magnitude of *trans*-acting regulatory components.

1.4.3 Epigenetic and Dosage Effects

Another dimension quantifying the molecular basis of heterosis involves non-coding RNA (ncRNA) and their epigenetic regulatory role in post-translational genome function via histone octamer modifications (Chen 2013). Most small interfering RNA (siRNAs) are derived from transposable and repeat elements and thus have diverged between species, in newly formed hybrids, their expression absent in the F_1 (but present in the parents) are restored in later generations. These differences in siRNA levels between hybrids or allopolyploids and their parents could alter allelic patterns of expression, RNA-directed DNA methylation (RdDM), and

overall genomic stability. Another ncRNA, micro RNA-encoding loci (miRNA) are transcribed by the RNA polymerase II complex and lead to the degradation or ubiquitination of target mRNAs through the effector RNA-induced silencing complex (RISC). In addition, miRNA precursors originating from different progenitors may have different processing efficiencies in polyploids (Ng et al., 2012). Though DNA methylation is one form of epigenetic mark that has been widely associated with gene expression changes, Donoghue et al. (2013) indicate that parental genome-dosage effects on gene expression in paternal-excess F_1 triploids are not associated with methylation and may instead be associated with RdDM pathways involving 24-nt small RNAs that are associated with *de novo* methylation. These findings are consistent with allelic bias of NBS-LRR transcript percentages in *Populus* F_1 hybrids compared to parental alleles with evidence of novel alternative splicing patterns. Donoghue et al. (2013) demonstrated novel parental dosage effects in isogenic maize F_1 triploids are methylation-independent at the whole-genome level, suggesting paternally and maternally inherited chromosome sets in autopolyploid plants may be epigenetically different, as a result of parental genome-dosage effects that can affect transcript levels in a methylation-independent manner. Despite the fact that the reciprocal F_1 triploids were genetically identical at the DNA sequence level, 602 genes were found to be differentially expressed between the reciprocal F_1 hybrids. Of these genes, all were up-regulated in the paternal-excess F_1 triploid compared with the maternal-excess F_1 triploid. Screening these genes for GO terms discovered a significant parent-of-origin genome dosage effect on stress response genes, with several response terms significantly overrepresented in the genes upregulated in the paternal F_1 triploid, including both biotic and abiotic stress response.

While the quantity of a gene product may impact the assembly of a particular complex, the mere involvement in a complex can also impact protein stability (Veitia et al., 2008). The

balance of regulatory hierarchies, i.e., the ‘dosage balance hypothesis’ (Birchler et al., 2005), are sensitive to gene dosage and changes in individual components can have an effect on the phenotype. In macromolecular complexes, dosage balance is essential because partial aneuploidy of a dosage-sensitive gene will change the abundances of the complexes and lead to fitness defects (Veitia et al., 2008). Contrary to the complementation theory of heterosis, it may be that null mutations in metabolic functions are tolerated in a heterozygous state, but only weak, loss-of-function, dosage-sensitive genes can survive negative selection as heterozygotes (Birchler and Veitia, 2010).

1.4.4 Other Technical Considerations

One of the most important factor in transcriptomics is building a quality whole genome assembly that is representative of a population, as opposed to fitting a sample population into a single reference. It is well-recognized that isoform data should be included in reference assemblies, but building quality *de novo* assemblies are technically difficult (computationally and biologically) and often avoided due to added filtering steps, fluctuations via small changes in assembly parameters, aberrant repeat elements, isoform detection, and the time-consuming annotation pipeline - all of which increase with sample size and ploidy-level. In well-known systems, like Arabidopsis, these analyses have been streamlined, considering the depth of collaborative input (e.g., TAIR, AthaMap, and Araport) in the past few decades. For instance, quality filtering paired-end reads, mapping to a reference genome or transcriptome, variant detection, merging annotations, and normalizing and quantifying count data can be accomplished with a laptop with a few CPU cores and 16 GB memory. Conversely, transcriptomics in non-model species must consider alternative approaches to attaining useful results.

Reference bias is often discussed in bioinformatics publications, but is only rarely accounted for. This reference bias exists because the contrasting splicing patterns between parents and progeny may be influenced by the direction of the reference template, so varying sequence identities (bp and % mapped), e.g., splicing multiple times within a gene, between tissue-type and individual and are thrown out by the assembler. For the reason that reads are generally smaller than transcripts from which they are derived, a single read may map to multiple gene isoforms, complicating expression analyses. However, an alternative way to approach isoform uncertainty is to explain gene expression in terms of the *sum* of isoform expression rather than read mapping, thus retaining both alternative isoform and valuable polymorphism which may be lost by mapping to the 'major' transcriptome. Likewise, by initially setting conservative mapping parameters, and then assembling *de novo* non-reference-mapped reads per sample, inclusion of the 'left-over subset' may be useful in comparing some of the anomalies that exist between samples and the reference. However, utilizing low-coverage DNA-Seq in coordination with RNA-Seq can improve confidence in calling RNA-Seq variants (which are often subject to allele-bias) from reference alignments (Carlson et al., 2017) and can even be used to build multiple parent references in interspecific gene expression studies.

By considering prior approximations of parental divergence of nuclear or cytoplasmic DNA within *de novo* assemblies, comparisons of duplicated genes (K_s) will give a useful estimates of under- or over-assembly, if such events are known. In addition, by including genetic information (whole chromosome) into transcriptome analyses, it is possible to account for a portion of the transcriptome variation between the dioecious hybrid and the parents by attributing sex (sex-chromosomes) or cytoplasmic (mitochondria and plastids) differences as model effects.

1.5 Evidence of Heterosis in Shrub Willow (*Salix* spp.)

1.5.1 Ecology and Uses of Willows

Shrub willow (*Salix* spp., Salicaceae) has been bred as an energy crop in Europe since the early 1970's with the goal of producing fast-growing bioenergy feedstock cultivars that are high-yielding, genetically diverse, pest and disease resistant, and able to grow on marginal land without competing with food crops. The genus *Salix* includes species that are naturally diploid, tetraploid, hexaploid, or with varying ploidy levels up to dodecaploid (Suda and Argus 1968; Thibault 1998). More than 300 species have been described in *Salix* (Argus 1997), some of which can readily hybridize. Willows have tremendous ecological amplitude – spread among marginal and riparian habitats from the arctic plains to the subtropics.

The particularly astringent phenolic glycosides (namely salicin) and other secondary metabolites found in willow bark provided early Europeans and Native American tribes an ethnobotanical remedy for a number of ailments and other utilitarian purposes. Although natural history museums are relatively deplete of Native American willow artifacts (Smart, personal communication), its utility is well-documented. The avid herbalist, Erichsen-Brown (1979), concatenates a number of historical references to willow in her landmark book "Medicinal and other Uses of North American Plants." For example, the Cheyenne were known to mix "the powdered rhizome and roots (*Acorus calamus*) with powdered red willow bark (*S. sericea*) and used the mixture for smoking" if tobacco was not available (91, ch. 2). Smith (1933) notes the Forest Potawatomi used the root of "shining willow" (*S. lucida*) as one of the ingredients of a remedy which was snuffed up the nose to stop hemorrhaging. Though the stem galls of willow had no use to the Patwatomi, the Menomini used them in tea and other remedies (94; ch. 2). The Hudson Bay natives broke willow shoots while hunting and were reported to "lay [willow]

across his track, by which the woman knows where to find such deer killed." The Hudson Bay tribe also constructed snowshoes and sleds from willow shoots during the winter months (92; ch. 2).

The number and heterogeneity of species and adaptive plasticity of willow has provided geneticists and breeders a great pool of germplasm for trait improvement and phylogenetics. Although breeding of willow as a bioenergy crop was initiated in Sweden and the United Kingdom in the early 1980's, the hybrids developed in the European breeding programs based on *S. viminalis* have been tested in North America and most are strongly debilitated by potato leaf hopper (*Empoasca fabae*) (Kopp 2000). Thus, regionally adapted cultivars for North America have been bred and selected from novel pedigrees, and corresponding mapping populations are continually being developed. Since 1998, willow breeding in NY has involved more than 750 controlled crosses utilizing a diverse collection of accessions representing over 25 species and their hybrids (Kopp et al., 2001; Smart et al., 2005; Smart and Cameron 2008; Smart et al., 2008; Smart and Cameron 2012).

1.5.2 Evidence of Heterosis in Interspecific Willow Crosses

Hybridization has been a primary approach in breeding shrub willow (*Salix* spp.) because hybrids often display heterosis for yield. Genetic improvement of willow has relied primarily on capturing heterosis through interspecific hybridization, and high-yielding elite cultivars have been selected, scaled-up, and commercialized (Double A Willow, Fredonia, NY). Increased vigor can be observed in intraspecific crosses, but the effect is more dramatic in interspecific hybrids, especially in triploid F_1 progeny derived from the hybridization of tetraploid and diploid parents. Many commercial cultivars are triploid hybrid F_1 progeny of the interspecific crosses *S. purpurea* (2n=2x=38) × *S. miyabeana* (4X) ('Millbrook' and 'Oneida') and *S. viminalis*

(2n=2x=38) × *S. miyabeana* (2n=2x=76) ('Preble', 'Fabius', 'Otisco', 'Owasco', and 'Tully Champion'). Outperforming foundation commercial cultivars, these hybrids show great promise for the future of the biomass production industry (Fabio et al. 2016).

It is vital that we apply our ever-improving understanding of heterosis from studies of well-characterized diploid crop species, such as maize, tomato, and rice to the improvement of yield and biomass quality of undomesticated crops, including willow and poplar that are being developed to provide sustainable sources of lignocellulosic biomass for bioenergy, biofuels, and bioproducts. As the challenge to meet global energy consumption increases along with global temperatures, the demand for low-input renewable bioproducts gives shrub willow great potential as a competitive bioenergy feedstock for biomass production and conversion to biofuels to offset fossil fuel usage. Significant effort is needed to develop accelerated crop breeding strategies and with the advent of high-throughput, next-generation sequencing technologies, and low-cost genotyping protocols, the ability to accelerate breeding and selection, especially with non-model crops, will continue to improve.

1.5.3 Current Status of Willow Genomics

Comparative genomic and trait analysis in willow will greatly improve opportunities for innovative manipulation of woody crops for improved feedstock production and will contribute to our understanding of the genomic basis for heterosis in diploid and polyploid interspecific hybrids in the Salicaceae. A necessary foundation for this research is a high-quality whole genome assembly and annotation of the *Salix* genome. My efforts have focused on *Salix purpurea* in the development of a reference genome for the genus. An Illumina-based pseudomolecule assembly (Figure 1.3) has been assembled and annotated with partners at JGI

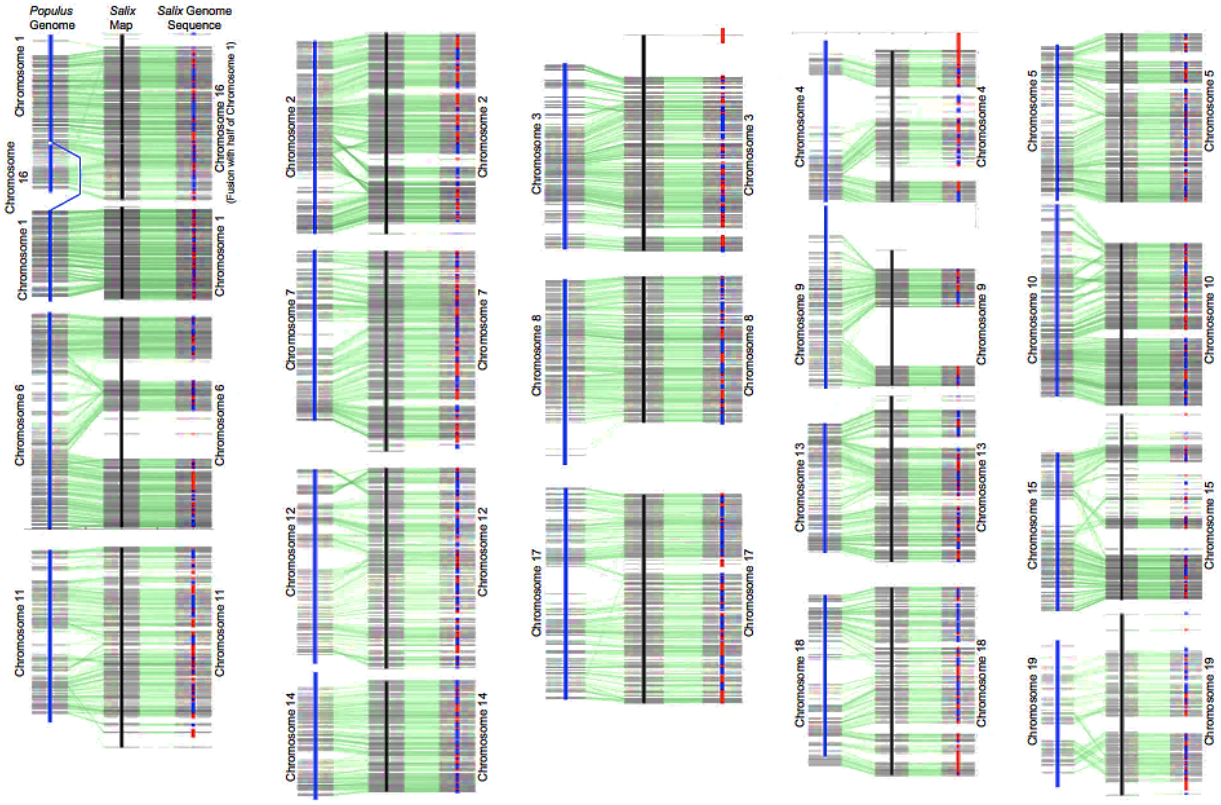

Figure 1.3 Pseudomolecules of the *S. purpurea* v1 reference genome anchored to the F_2 intercross linkage map (see Chapter 4) and aligned with the *Populus trichocarpa* v3 reference genome, to the right, center, and left of each chromosomal alignment, respectively. Figure courtesy of Eli Rodgers-Melnick and Stephen DiFazio.

and JCVI (*Salix purpurea* v1.0, phytozome.jgi.doe.gov). Since the first *Salix* genome assembly, the willow genomics group has worked on the development of separate male and female assemblies of *S. purpurea*, using a long read sequencing platform, and are currently integrating information from the two assemblies to resolve haplotypes collapsed in the assembly process, which not uncommon for heterozygous species.

1.6 Main Questions

One of the major challenges in molecular genetics is disentangling the relationship of transcriptome-wide expression patterns to phenotypic effects (Birchler 2007). While there have

been many studies on differential gene expression in intra- and interspecific hybrids, most were generated from crossing inbred parents; hence, the effects of intra- and interspecific hybridization on gene expression in polyploid outcrossing populations of dioecious perennial species are largely unknown. There are publications which do include the genomic and transcriptomic effects of neo-polyploid formation, yet its treatment in the literature is a broad stroke when applied to plants that do not fit an inbred diploid model and accompanies only speculation. Whether triploid heterosis is a result of genetic or macromolecular factors, parental divergence is undoubtedly a substantial factor and has been realized and applied in a plethora of breeding programs.

The main focus of this study was to explore the regulatory architectures of gene expression in intra- and interspecific diploid and triploid hybrids of shrub willow and to identify genes responsible for heterosis, especially for those genes related to biomass (Osborn et al. 2003; Ni et al. 2009; Stokes et al. 2010; Goff 2011) or wood composition quality (Serapiglia et al. 2009; Serapiglia et al. 2012), that correlate with yield heterosis in the progeny. A set of eight families were planted in the greenhouse and field to evaluate phenotypic heterosis for a suite of important biomass and wood composition traits, then examined for patterns of expression within and among hybrid category suggestive of *cis*-acting bias or *trans*-acting multigenic control.

The primary questions this study has sought to answer are as follows: 1) Do triploid hybrids outperform their progenitors for important biomass-related traits? 2) To what extent is gene expression inheritance non-additive? 3) Is there allele-specific expression inherited from the parents of species hybrids? 4) Is there a dosage effect on the expression of parent alleles in the triploid hybrid? 5) Are differentially-expressed genes responsible for heterosis for important biomass and wood composition traits?

1.7 REFERENCES

Argus GW. 1997. Infrageneric Classification of *Salix* (Salicaceae) in the New World. Systematic Botany Monographs. Laramie, WY, American Society of Plant Taxonomists. 52: 1–121.

Auger DL, Gray AD, Ream TS, Kato A, Coe EH, Birchler JA. 2005. Nonadditive gene expression in diploid and triploid hybrids of maize. *Genetics*, 169: 389–397.

Benson MK, Einspahr DW. 1967. Early growth of diploid triploid and triploid hybrid aspen. *Forest Science*, 13: 150.

Bergstrom I. 1940. On the progeny of diploid × triploid *Populus tremula* with special reference to the occurrence of tetraploidy. *Hereditas*, 26: 191–201.

Bingham ET, Groose RW, Woodfield DR, Kidwell KK. 1994. Complementary gene interactions in alfalfa are greater in autotetraploids than diploids. *Crop Science*, 34: 823–829.

Birchler JA, Auger DL, Riddle NC. 2003. In search of the molecular basis of heterosis. *The Plant Cell*, 15: 2236–2239.

Birchler JA, Riddle NC, Auger DL, Veitia RA. 2005. Dosage balance in gene regulation: biological implications. *Trends in Genetics*, 21: 219–226.

Birchler JA, Veitia RA. 2007. The gene balance hypothesis: From classical genetics to modern genomics. *The Plant Cell*, 19: 395–402.

Birchler, JA. 2010. Heterosis. *The Plant Cell*, 22: 2105–2112.

Bradshaw HD Jr., Stettler RF. 1993. Molecular genetics of growth and development in *Populus*. I. Triploidy in hybrid poplars. *Theoretical and Applied Genetics*, 86: 301–307.

Bradshaw HD, Stettler RF. 1995. Molecular genetics of growth and development in Populus. IV. Mapping QTLs with large effects on growth, form, and phenology traits in a forest tree. *Genetics*, 139: 963–973.

Carlson CH, Choi Y, Chan AP, Serapiglia MJ, Town CD, Smart LB. 2017. Dominance and sexual dimorphism pervade the *Salix purpurea* L. transcriptome. *Genome Biology and Evolution*, 9: 2377–2394.

Casler MD, Stendal CA, Kapich L, Vogel KP. 2007. Genetic diversity, plant adaptation regions, and gene pools for switchgrass. *Crop Science*, 47: 2261–2273.

Charlesworth D, Willis JH. 2009. The genetics of inbreeding depression. *Nature Reviews Genetics*, 10: 783–796.

Clifton-Brown JC, Lewandowski I. 2002. Screening *Miscanthus* genotypes in field trials to optimise biomass yield and quality in Southern Germany. *European Journal of Agronomy*, 16: 97–110.

Darwin C. 1876. The effects of cross fertilization in the vegetable kingdom. J. Murray, 1876.

Duvick DN. 1999. Heterosis: Feeding People and Protecting Natural Resources. In: J. G. Coors and S. Pandey (Eds.) The Genetics and Exploitation of Heterosis in Crops. Madison, WI, American Society of Agronomy; Crop Science Society of America; Soil Science Society of America: 19–29.

East EM. 1936. Heterosis. *Genetics*, 21: 375–397.

Einspahr DW, Benson MK, Peckham JR. 1968. Wood and pulp properties of 5-year-old diploid triploid and triploid hybrid aspen. *TAPPI Journal*, 51: 72.

Einspahr DW, Peckham JR, Benson MK. 1970. Fiber and pulp properties of triploid and tetraploid hybrid aspen. *TAPPI Journal*, 53: 1853.

Einspahr DW, van Buijtenen JP, Peckham JR. 1963. Natural variation and heritability in triploid Aspen. *Silvae Genetica*, 12: 51–58.

Erichsen-Brown C. 1979. Medicinal and other uses of North American plants: a historical survey with special reference to the Eastern Indian tribes. Dover Publications, Inc. Mineola, NY.

Fabio ES, Volk TA, Miller RO, Serapiglia MJ, Gauch HG, Van Rees KCJ, Hangs RD, Amichev BY, Kuzovkina YA, Labrecque M, Johnson GA, Ewy RG, Kling GJ, Smart LB. 2016. Genotype × environment interaction analysis of North American shrub willow yield trials confirms superior performance of triploid hybrids. *GCB Bioenergy*, 9: 445–459.

Fu H, Dooner HK. 2002. Intraspecific violation of genetic colinearity and its implications in maize. *Proceedings of the National Academy of Sciences U. S. A.*, 99: 9573-78.

Glowacka K, Jezowski S, Kaczmarek Z. 2010. In vitro induction of polyploidy by colchicine treatment of shoots and preliminary characterisation of induced polyploids in two *Miscanthus* species. *Industrial Crops and Products*, 32: 88–96.

Goff SA. 2011. A unifying theory for general multigenic heterosis: energy efficiency, protein metabolism, and implications for molecular breeding. *New Phytologist*, 189: 923–37.

Guo M, Davis D, Birchler JA. 1996. Dosage effects on gene expression in a maize ploidy series. *Genetics*, 142: 1349–1355.

Guo M, Rupe MA, Yang X, Crasta O, Zinselmeier C, Smith OS, Bowen B. 2006. Genome-wide transcript analysis of maize hybrids: allelic additive gene expression and yield heterosis. *Theoretical and Applied Genetics*, 113: 831–845.

Guo M, Rupe MA, Zinselmeier C, Habben J, Bowen BA, Smith OS. 2004. Allelic variation of gene expression in maize hybrids. *The Plant Cell*, 16: 1707–1716.

Hardig TM, Brunsfeld SJ, Fritz RS Morgan M, Orians S. 2000. Morphological and molecular evidence for hybridization and introgression in a willow (*Salix*) hybrid zone. *Molecular Ecology*, 9: 9–24.

He GM, Zhu XP, Elling AA, Chen LB, Wang XF, Guo L, Liang MZ, He H, Zhang HY, Chen FF, Qi YJ, Chen RS, Deng, XW. 2010. Global epigenetic and transcriptional trends among two rice subspecies and their reciprocal hybrids. *The Plant Cell*, 22: 17–33.

Hornsey KG. 1975. The exploitation of polyploidy in sugar beet breeding. *Journal of Agricultural Science*, 84: 543–558.

Hua J, Xing Y, Wu W, Xu C, Sun X, Yu S, Zhang Q. 2003. Single-locus heterotic effects and dominance by dominance interactions can adequately explain the genetic basis for heterosis in an elite rice hybrid. *Proceedings of the National Academy of Sciences U. S. A.*, 100: 2574–2579.

Huang X, Yang S, Gong J, Zhao Q, Feng Q, Zhan Q, Zhao Y, Li W, Cheng B, Xia J, Chen N, Huang T, Zhang L, Fan D, Chen J, Zhou C, Lu Y, Weng Q, Han B. Genomic architecture of heterosis for yield traits in rice. *Nature*, 537: 629–633.

Kelleher CT, Chiu R, Shin H, Bosdet IE, Krzywinski MI, Fjell CD, Wilkin J, Yin TM, DiFazio SP, Ali J, Asano JK, Chan S, Cloutier A, Girn N, Leach S, Lee D, Mathewson CA, Olson T, O'Connor K, Prabhu AL, Smailus DE, Stott JM, Tsai M, Wye NH, Yang GS, Zhuang J, Holt RA, Putnam NH, Vrebalov J, Giovannoni JJ, Grimwood J, Schmutz J, Rokhsar D, Jones SJM, Marra MA, Tuskan GA, Bohlmann J, Ellis BE, Ritland K, Douglas CJ, Schein JE. 2007. A physical map of the highly heterozygous *Populus* genome: integration with the genome sequence and genetic map and analysis of haplotype variation. *The Plant Journal*, 50: 1063–1078.

Kopp RF. 2000. Genetic improvement of *Salix* using traditional breeding and AFLP fingerprinting. Thesis. SUNY College of Environmental Science and Forestry, Syracuse, NY. pp. 175.

Kopp RF, Smart LB, Maynard CA, Isebrands JG, Tuskan GA, Abrahamson LP. 2001. The development of improved willow clones for eastern North America. *Forestry Chronical*, 77: 287–292.

Kopp RF, Smart LB, Maynard CA, Tuskan GA, Abrahamson LP. 2002. Predicting within-family variability in juvenile height growth of *Salix* based upon similarity among parental AFLP fingerprints. *Theoretical and Applied Genetics*, 105: 106–112.

Krieger U, Lippman ZB, Zamir D. 2010. The flowering gene SINGLE FLOWER TRUSS drives heterosis for yield in tomato. *Nature Genetics*, 42: 459–463.

Larièpe A, Mangin B, Jasson S, Combes V, Dumas F, Jamin P, Lariagon C, Jolivot D, Madur D, Fiévet J, Gallais A, Dubreuil P, Charcosset A, Moreau L. 2012. The genetic basis of heterosis: Multiparental quantitative trait loci mapping reveals contrasted levels of apparent overdominance among traits of agronomical interest in maize (*Zea mays* L.). *Genetics*, 190: 795–811.

Li B, Wu R. 1996. Genetic causes of heterosis in juvenile aspen: A quantitative comparison across intra- and inter-specific hybrids. *Theoretical and Applied Genetics*, 93: 380–391.

Lin J, Gibbs JP, Smart LB. 2009. Population genetic structure of native versus naturalized sympatric shrub willows (*Salix*; Salicaceae). *American Journal of Botany*, 96: 771-85.

Lippman ZB, Zamir D. 2007. Heterosis: revisiting the magic. *Trends in Genetics*, 23: 60-66.

Marron N, Ceulemans R. 2006. Genetic variation of leaf traits related to productivity in a *Populus deltoides* × *Populus nigra* family. *Canadian Journal of Forest Research,* 36: 390–400.

McManus CJ, Coolon JD, Duff MO, Eipper-Mains J, Graveley BR, Wittkopp PJ. 2010. Regulatory divergence in Drosophila revealed by mRNA-seq. *Genome Research,* 20: 816–825.

Meiklejohn CD, Coolon JD, Hartl DL, Wittkopp PJ. 2014. The roles of *cis*- and *trans*-regulation in the evolution of regulatory incompatibilities and sexually dimorphic gene expression. *Genome Research*, 24: 84–95.

Müntzing A. 1936. The chromosomes of a giant *Populus tremula*. *Hereditas,* 21: 383-93.

Ni ZF, Kim ED, Ha MS, Lackey E, Liu JX, Zhang YR, Sun QX, Chen ZJ. 2009. Altered circadian rhythms regulate growth vigor in hybrids and allopolyploids. *Nature*, 457: 327.

Nilsson-Ehle H. 1936. Über eine in der natur gefundene gigasform von *Populus tremula*. *Hereditas,* 21: 379–382.

Osborn TC, Pires JC, Birchler JA, Auger DL, Chen ZJ, Lee HS, Comai L, Madlung A, Doerge RW, Colot V, Martienssen RA. 2003. Understanding mechanisms of novel gene expression in polyploids. *Trends in Genetics,* 19: 141–147.

Rédei GP. 1962. Single locus heterosis. *Zeitschrift für Vererbungslehre*, 93: 164-170.

Riddle N, Birchler J. 2008. Comparative analysis of inbred and hybrid maize at the diploid and tetraploid levels. *Theoretical and Applied Genetics*, 116: 563–576.

Riddle N, Kato A, Birchler J. 2006. Genetic variation for the response to ploidy change in Zea mays L. *Theoretical and Applied Genetics,* 114: 101–111.

Riddle NC, Jiang HM, An LL, Doerge RW, Birchler JA. 2010. Gene expression analysis at the intersection of ploidy and hybridity in maize. *Theoretical and Applied Genetics,* 120: 341–353.

Robinson JC, Saúco VG. 2010. Bananas and Plantains, 2nd ed. Wallingford, UK, CABI Publishing. DOI: 10.1079/9781845936587.0000.

Rodgers-Melnick E, Mane SP, Dharmawardhana P, Slavov GT, Crasta OR, Strauss SH, Brunner AM, DiFazio SP. 2012. Contrasting patterns of evolution following whole genome versus tandem duplication events in *Populus*. *Genome Research*, 22: 95–105.

Savitsky VF. 1962. Sucrose and weight of root in tetraploid monogerm and multigerm sugar beet populations under different mating systems. *Journal of the American Society of Sugar Beet*, 11: 676–711.

Scascitelli M, Cognet M, Adams KL. 2010. An interspecific plant hybrid shows novel changes in parental splice forms of genes for splicing factors. *Genetics*, 184: 975–U136.

Serapiglia MJ, Cameron KD, Stipanovic AJ, Smart LB. 2009. Analysis of biomass composition using high-resolution thermogravimetric analysis and percent bark content for the selection of shrub willow bioenergy crop varieties. *Bioenergy Research*, 2: 1–9.

Serapiglia MJ, Cameron KD, Stipanovic AJ, Smart LB. 2012. Correlations of expression of cell wall biosynthesis genes with variation in biomass composition in shrub willow (*Salix* spp.) biomass crops. *Tree Genetics and Genom*es, 8: 775–788.

Shi JQ, Li RY, Zou J, Long Y, Meng JL. 2011. A dynamic and complex network regulates the heterosis of yield-correlated traits in rapeseed (*Brassica napus* L.). *PLoS ONE*, 6: e21645.

Shull GH. 1946. Hybrid Seed Corn. *Science*, 103: 547–550.

Smart LB, Cameron KD. 2008. Genetic improvement of willow (*Salix* spp.) as a dedicated bioenergy crop. In: W. E. Vermerris (Ed.) Genetic Improvement of Bioenergy Crops. New York, NY, Springer Science: 347–376.

Smart LB, Cameron KD. 2012. Shrub willow (*Salix* spp.). In: C. Kole, S. Joshi and D. Shonnard (Eds.), Handbook of Bioenergy Crop Plants. Boca Raton, FL, Taylor and Francis Group, pp. 687–708.

Smart LB, Cameron KD, Volk TA, Abrahamson LP. 2008. Breeding, selection, and testing of shrub willow as a dedicated energy crop. In: A. Eaglesham (Ed.), NABC Report 19 – Agricultural Biofuels: Technology, Sustainability and Profitability. Ithaca, NY, National Agricultural Biotechnology Council.

Smart LB, Volk TA, Lin J, Kopp RF, Phillips IS, Cameron KD, White EH, Abrahamson LP. 2005. Genetic improvement of shrub willow (*Salix* spp.) crops for bioenergy and environmental applications in the United States. *Unasylva*, 56: 51–55.

Smith HH. 1933. Ethnobotany of the Forest Potawatomi Indians. Bulletin of the Public Museum of the City of Milwaukee, 7: 1–230, pp. 56–57, 80–81.

Springer NM, Stupar RM. 2007a. Allele-specific expression patterns reveal biases and embryo-specific parent-of-origin effects in hybrid maize. *The Plant Cell*, 19: 2391–2402.

Springer NM, Stupar RM. 2007b. Allelic variation and heterosis in maize: How do two halves make more than a whole? *Genome Research*, 17: 264–275.

Sreekumari MT, Jos JS, Nair SG. 1999. Sree Harsha: A superior triploid hybrid in cassava. *Euphytica*, 106: 1–6.

Stettler RF, Zsuffa L, Wu R. 1996. The role of hybridization in the genetic manipulation of Populus. In: R. F. Stettler, H. D. Bradshaw, P. E. Heilman and T. M. Hinckley (Eds.)

Biology of *Populus* and its Implications for Management and Conservation. Ottawa, Ontario, Canada, NRC Research Press: 87–112.

Stokes D, Fraser F, Morgan C, O'Neill CM, Dreos R, Magusin A, Szalma S, Bancroft I. 2010. An association transcriptomics approach to the prediction of hybrid performance. *Molecular Breeding*, 26: 91–106.

Stupar RM, Gardiner JM, Oldre AG, Haun WJ, Chandler VL, Springer NM. 2008. Gene expression analyses in maize inbreds and hybrids with varying levels of heterosis. *BMC Plant Biology*, 8: 33.

Stupar RM, Springer NM. 2006. *Cis*-transcriptional variation in maize inbred lines B73 and Mo17 leads to additive expression patterns in the F_1 hybrid. *Genetics*, 173: 2199-210.

Suda Y, Argu GW. 1968. Chromosome numbers of some North American *Salix*. *Brittonia*, 20: 191–197.

Swanson-Wagner RA, Jia Y, DeCook R, Borsuk LA, Nettleton D, Schnable PS. 2006. All possible modes of gene action are observed in a global comparison of gene expression in a maize F_1 hybrid and its inbred parents. *Proceedings of the National Academy of Sciences U. S. A.*, 103: 6805–6510.

Tanksley SD. 1993. Mapping polygenes. *Annual Reviews Genetics*, 27: 205–233.

Thibault J. 1998. Nuclear DNA amount in pure species and hybrid willows (*Salix*): a flow cytometric investigation. *Canadian Journal of Botany*, 76: 157–165.

Veitia RA, Bottani S, Birchler JA. 2008. Cellular reactions to gene dosage balance: Genomic, transcriptomic and proteomic effects. *Trends in Genetics*, 24: 390–397.

Vrebalov J, Ruezinsky D, Padmanabhan V, White R, Medrano D, Drake R, Schuch W, Giovannoni J. 2002. A MADS-box gene necessary for fruit ripening at the tomato ripening inhibitor (*rin*) locus. *Science*, 343–346.

Wei G, Tao Y, Liu GZ, Chen C, Luo RY, Xia HA, Gan Q, Zeng HP, Lu ZK, Han YN, Li XB, Song GS, Zhai HL, Peng YG, Li DY, Xu HL, Wei XL, Cao ML, Deng HF, Xin YY, Fu XQ, Yuan LP, Yu J, Zhu Z, Zhu LH. 2009. A transcriptomic analysis of superhybrid rice LYP9 and its parents. *Proceedings of the National Academy of Sciences U. S. A.*, 106: 7695–7701.

Wittkopp PJ, Haerum BK, Clark AG. 2008. Independent effects of *cis*- and *trans*-regulatory variation on gene expression in *Drosophila melanogaster*. *Genetics*, 178: 1831–1835.

Wittkopp PJ, Kalay G. 2012. *Cis*-regulatory elements: Molecular mechanisms and evolutionary processes underlying divergence. *Nature Review Genetics*, 13: 59–69.

Winton L, Einspahr DW. 1970. Tetraploid aspen production using unreduced triploid pollen. *Forest Science*, 16: 180–182.

Wray G. 2007. Evolutionary significance of *cis*-regulatory mutations. *Nature Review Genetics*, 8: 206–216.

Zamir D. 2001. Improving plant breeding with exotic genetic libraries. *Nature Review Genetics*, 2: 983–989.

Zhuang Y, Adams KL. 2007. Extensive allelic variation in gene expression in *Populus* F_1 hybrids. *Genetics*, 177: 1987–1996.

Zsuffa L, Mosseler A, Raj Y. 1984. Prospects for interspecific hybridization in willow for biomass production. In: K. Perttu (Ed.) Ecology and Management of Forest Biomass Production Systems. Uppsala, Sweden, Swedish University of Agricultural Sciences.

CHAPTER 2

ELECTRICAL CAPACITANCE AS A PREDICTOR OF ROOT DRY WEIGHT IN SHRUB WILLOW (*SALIX*; SALICACEAE) PARENTS AND PROGENY

Published as: Carlson CH, Smart LB. 2016. *Applications in Plant Science*, 4(8): e1600031.

2.1 Abstract

Root biomass is an important trait often disregarded in woody perennial selection due to the challenge and expense of accurately and efficiently measuring large populations. In this study, we aim to develop a simple method that can predict root dry weight within a diverse shrub willow (*Salix*) breeding population representing species hybrids and their parents using root electrical capacitance (REC). The REC method was tested on plants started from cuttings and grown in pots with potting mix in the greenhouse for 11 wk to assess the relationship of REC with 24 biomass traits and its usefulness in allometric models for root and stem dry biomass. Strong linear and positive correlations were found between REC and root dry biomass (r^2 = 0.88). The total proportion of variance of root and stem dry biomass explained by predictors in multiple regression was 85% and 69%, respectively. The relative importance of predictor variables in allometric models were dominated by the contribution of REC. This work provides an efficient and non-destructive technique to indirectly quantify root biomass of genetically diverse shrub willow progeny, which has great promise for selection of genotypes with varying root biomass and for the accurate estimation of belowground carbon sequestration.

2.2 Introduction

The measurement of belowground traits of woody perennials has been the subject of renewed interest in recent years, since root biomass and architecture is critical for drought tolerance and for long-term carbon sequestration. Measurement of root biomass is a tremendously difficult and painstaking task in woody perennial breeding programs, as it requires careful washing, filtering, drying, and weighing fine root tissues, and thus is often avoided. For decades, electrical capacitance has been used to estimate the aboveground biomass of herbaceous perennials (Currie et al., 1987) and forest plantations (Lekas et al., 1990) but few root electrical capacitance (REC) studies have focused on clonal woody perennial plants as a model for predicting root biomass. Previous work on the estimation of root biomass using the REC method have primarily focused on hydroponically-grown seedlings (Rajkai et al., 2005; Repo et al., 2005; Cao et al., 2011; Dietrich et al., 2012; Cseresnyés et al., 2013; Kormanek and Tomasz, 2015) or a limited number of genotypes (Whitlow et al., 1992; Pitre et al., 2010).

Electrical capacitance is measured in farads (F) and is considered the ability of a plant tissue (roots) to store an electrical charge, such that when charged with one coulomb (C) of energy, there is a potential difference of one volt (V), where F = C/V. Dalton (1995) provided the first and most widely-accepted conceptual model for using electrical capacitance as an *in situ* measurement for assessing root development. His method assumes that root capacitance is equivalent to a parallel resistance-capacitance circuit formed by the interface between soil and water and the root surface; or simply, that roots are equivalent to cylindrical capacitors where the epidermis and xylem are external and internal electrodes, respectively.

Using hydroponically-grown barley seedlings, Dietrich et al. (2012) argued that capacitance is linearly correlated with the sum root cross-sectional areas at the solution surface

and inversely related to the distance between the plant electrode and the solution surface. Likewise, the capacitance of barley is not determined by actual root biomass, but by the cross-sectional area of roots at the solution surface (Dietrich et al., 2013). The distinction between this model and Dalton's is that Dalton generalized roots to be cylindrical capacitors acting in parallel, whereas Dietrich argued that the capacitances of tissues along an unbranched root can be considered to be connected in series and the entire root system, in parallel.

Although there was a strong linear relationship between the cross-sectional area of the stem and the resistance of the stem, Repo et al. (2005) found that there was no difference in the relationship between resistance and cross-sectional area with or without the roots attached. Ellis et al. (2013) found little evidence capacitance alone could be related to root mass but using these measurements along with approximations of average root tissue fresh density could be used together to estimate coarse root length. Dietrich et al. (2012) provides good evidence to modify the conceptual framework of Dalton's original model, yet all analyses were based on fresh root mass of a hydroponically-grown herbaceous annual. Compared to dry mass, wet mass can be extremely variable in woody plants and heavily dependent on environmental factors, and is not an ideal estimate of root dry weight biomass. Further, the compositional differences and developmental stage between seedling and mature, fully-developed root systems cannot be ignored. Here, we simply assume the Dalton model using a diverse breeding population of shrub willow (*Salix*; Salicaceae) with no assumptions concerning root morphology or architecture.

Shrub willow has been bred as an energy crop in Europe since the early 1970's with the goal of producing fast-growing bioenergy feedstock cultivars that are high-yielding, genetically diverse, pest and disease resistant, and able to grow on marginal land without competing with food crops (Karp et al. 2011; Smart and Cameron, 2012). Shrub willow are dioecious riparian

species commonly found near riverbanks and streambeds and are well adapted to the hypoxic conditions of wetlands (Kuzovkina et al., 2008). More than 350 species have been described in *Salix* (Argus, 2005; Lauron-Moreau et al., 2015) and have tremendous ecological amplitude, found within marginal and riparian habitats from the arctic plains to the subtropics (Kuzovkina et al., 2008). As the challenge to meet global energy objectives increases along with global temperatures (Walther et al., 2002; Whiteman et al., 2013), the demand for low-input renewable bioproducts gives shrub willow great potential as a competitive bioenergy feedstock for biomass production and conversion to biofuels to offset fossil fuel usage (Bonosi et al., 2013).

Our main objectives of this study were: (1) to investigate the relationship between electrical capacitance and root dry weight in diverse shrub willow families; (2) to contrast REC of parent genotypes and their hybrid progeny; (3) to study the relationship between REC and above- and belowground biomass traits; and (4) to evaluate the relative importance of REC in allometric models for the prediction of root and stem dry weight.

2.3 Materials and Methods

2.3.1 Plant Material

Parent genotypes and randomly chosen progeny of *Salix purpurea* (2n=2x=38), *S. miyabeana* (2n=4x=76), and *S. viminalis* (2n=2x=38) species crosses were grown from stem cuttings (20 cm) in 12-L plastic pots with peat moss-based potting mix (Fafard, Agawam, MA) to evaluate growth traits under greenhouse conditions over 11 weeks. Families consisted of 12 progeny individuals and their parents from diploid intraspecific *S. purpurea* F_1 and F_2 crosses, two reciprocal diploid interspecific crosses (*S. purpurea* × *S. viminalis*), three triploid interspecific crosses (*S. purpurea* × *S. miyabeana*, *S. miyabeana* × *S. viminalis*, and *S. viminalis* × *S. miyabeana*), and intraspecific tetraploid cross of *S. miyabeana* for a total of 104 genotypes

(Table 2.1). One exception is that one of the parents of the *S. miyabeana* family (SX64, male parent of 425) was not included.

Table 2.1 Description of family, pedigree, and generation in intraspecific and interspecific crosses of shrub willow (*Salix*; Salicaceae).

Family ID	Female parent	Male Parent	Pedigree	F_n	Ploidy
082	94006	94001	*S. purpurea* × *S. purpurea*	F_1	2X
317	'Wolcott'	'Fish Creek'	*S. purpurea* × *S. purpurea*	F_2	2X
407	94006	'Jorr'	*S. purpurea* × *S. viminalis*	F_1	2X
421	07-MBG-5027	94001	*S. viminalis* × *S. purpurea*	F_1	2X
415	94006	01-200-003	*S. purpurea* × *S. miyabeana*	F_1	3X
423	07-MBG-5027	01-200-003	*S. viminalis* × *S. miyabeana*	F_1	3X
430	01-200-006	'Jorr'	*S. miyabeana* × *S. viminalis*	F_1	3X
425	01-200-006	SX64	*S. miyabeana* × *S. miyabeana*	F_1	4X

Plot was defined as a single cutting planted in a pot, which were arranged in a randomized complete block design with four replicate blocks. Two blocks were located on benches in one greenhouse with the other two blocks in an adjacent greenhouse set for identical growing conditions. Supplemental greenhouse lighting was provided on a 14 h day:10 h night regimen with max daytime temperature of 26°C and a nighttime temperature of 18°C. Beyond weekly applications of beneficial insects and mites for pest management, no pesticides were required, as there were no symptoms of biotic or abiotic stress on any plant material throughout the length of the study. Liquid fertilizer (Peter's 15-16-17 Peat-Lite Special®, Scott's, Marysville, OH) was applied weekly after week four according to manufacturer recommendations. The age of plants and the frequency of trait measurements is depicted in Table 2.2.

Table 2.2 List of biomass-related variables including their descriptions and units.

Trait Abbv.	Description	Units	Measured (WAP)[a]
Above ground			
SDW	Stem dry weight	g plant^{-1}	11
SSL	Sum stem length	cm	2–11
MSL	Mean stem length	cm	2–11
MXL	Max stem length	cm	2–11
MSA	Mean stem area	cm^2	11
SSA	Sum stem area	cm^2	11
RGR	Relative growth rate	cm d^{-1}	3–11
PSN	Primary stem number	#	11
ASN	Axial stem number	#	5, 10
LDW	Leaf dry weight	g plant^{-1}	11
SPD	SPAD	SPAD units	2, 4, 10, 11
PHE	Vegetative phenology	0→5	2–3
MSD	Mean stem diameter	cm	11
SSD	Sum stem diameter	cm	11
LFA	Leaf area	cm^2	11
SLA	Specific leaf area	cm^2 g^{-1}	10
LMA	Leaf mass per area	g cm^{-2}	10
Below ground			
REC	Root electrical capacitance	nF	11
RDW	Root dry weight	g plant^{-1}	11
CDIA	Cutting diameter	cm	11
CDW	Cutting dry weight	g plant^{-1}	11
RDIA	Mean root diameter	cm	11
PRN	Primary root number	#	11
SRN	Secondary root number	#	11
TRN	Total root number	#	11

[a]WAP, weeks after planting.

2.3.2 Root Electrical Capacitance

Root electrical capacitance was measured using a DCM3 digital capacitance meter (UEI Test Instruments™, Beaverton, OR) with conductance ranges of 200 pF to 20 mF, and a ±1% reading and digital accuracy below 2000 µF. Custom designed electrodes were attached to the capacitance meter as follows: the positive lead wire was connected to a 10 gauge (0.27 cm × 20 cm) solid copper wire, while the negative lead was connected to a 7.5 × 1.0 cm copper-plated gator clip. The copper wire attached to the positive lead of the EC meter was inserted 15 cm into the potting mix approximately 10 cm from the cutting and the negative lead was clamped to the cutting 2 cm above the surface of the potting mix. Root electrical capacitance was measured approximately 2 h after pots were completely saturated with reverse-osmosis water immediately after harvesting aboveground biomass. To minimize within-plot variation, four readings were taken equidistant to the cutting and then averaged for each pot. The appropriate experimental nominal capacitance was determined to be 200 nF or 20 µF (if EC >200 nF) at 820 and 82 Hz test frequencies, respectively.

2.3.3 Aboveground Biomass Traits

Starting approximately 7 d after planting, the vegetative phenology stage (PHE) of each plot was scored on five occasions over the next 8 d. Vegetative phenology was scored as six stages described as: stage (0) dormant axillary buds are tightly closed and covered by bud scales; (1) axillary buds begin to swell and change color; (2) generative bud burst with visible leaves; (3) leaves emerge and begin to unfold; (4) unfolded leaves begin expanding; and (5) at least two leaves are fully expanded.

Primary stems were defined as those emerging from dormant axillary buds and ≥6 cm in length. Secondary stems were defined as emerging from axillary buds on the primary stem from

current-season growth (sylleptic) and were counted as the total number of secondary branches within each plot with a vegetative phenology stage ≥ 3. The length of all stems ≥ 6 cm per plant was measured from the base of the primary stems at the original cutting surface to the distal shoot tip (the inner-whorl of the leaf primordia). Starting in the second week of the study until its termination, all primary stems were measured within each plot once a week over a period of 9 weeks.

Mean stem length (MSL) per plant was the mean of individual stem lengths and sum of stem length (SSL) was the sum of all primary stems measured for a plant. The diameter of each primary stem within a plot was measured at the base in the final week of the study using a digital caliper. The stem diameters were used to calculate mean stem area (MSA) for all of the stems on a plant, and all stem areas were summed to determine sum of stem area per plant (SSA). Only primary stems > 20 cm in length were used to calculate stem traits. Stem and leaf biomass were harvested separately from each plot, dried in an oven to constant weight at 65°C, and weighed to determine stem dry weight (SDW) and leaf dry weight (LDW).

Leaf area (LFA) was determined using a portable leaf area meter (Model No. CI-203, CID Inc., Camas, WA). A representative leaf from each plot was scanned, then excised, dried to constant weight at 65°C, and weighed to obtain leaf dry weight. Petioles were excluded from leaf area and dry weight measurements. Specific leaf area (SLA) was calculated as the ratio of the total leaf area and dry weight as well as its inverse, leaf mass per area (LMA). As an indirect measurement of leaf chlorophyll and nitrogen content, leaves were measured using a SPAD 502 Chlorophyll Meter (Spectrum Technologies, Inc., Aurora, IL). Four fully-expanded leaves from the upper 25% of the canopy were measured and averaged for each plant at weeks 2, 4, 6, and 10.

2.3.4 Belowground Biomass Traits

Root biomass was harvested from a subset of 20 pots (2-3 individuals per family) that were selected from the 416 pots as representing a distribution of capacitance readings, ranging from 70.5 nF (clone 13X-425-110) to 283.8 nF (clone 13X-430-033). To assure the retrieval of fine root hairs, potting mix was washed from roots by first soaking root balls in water for 12 h without their pots then rinsing them by hand repeatedly and decanting into 2.4 mm and 1.0 mm aluminum test sieves. Root samples were considered appropriate for dry weight analysis when root biomass was visually free of debris. Root biomass and cuttings were separately dried in an oven and weighed to obtain root dry weight (RDW) and cutting dry weight (CDW). The base diameter and number of each primary root as well as the average distance between secondary root primordia of primary roots were measured using an Absolute Digimatic® digital caliper (Model No. CD-6”CSX, Mitutoyo USA, Inc., Aurora, IL).

2.3.5 Statistical Analysis

All statistical analyses were performed in the statistical computing environment, R 3.2.2 (R Development Core Team, 2011). Traits (Table 1.2) were tested for paired associations ($\alpha = 0.05$) with REC using the cor.test function. To determine the best transformation of REC for the response variables, the Box-Cox procedure was used. Only SDW and LDW were found to be log-linear ($\lambda \sim 0$) to REC; consequently, a log transformation of REC was performed in linear and multiple regressions for these traits. When significant differences were observed, Tukey’s studentized range test was used for family mean comparisons and significant differences. Cross-validation of REC and RDW was accomplished via *K*-fold ($k = 5$) and leave-one-out (N - 1) methods using the least-square estimates of model coefficients as well as the root mean squared error (RMSE) to determine the prediction loss of the model in the package cvTools (Alfons,

2012). Variance components were estimated using lmer in the package lme4 (Bates, 2015) using the REML method. In order to analyze the effects of male, female, and male × female interactions, the following linear mixed model was used:

$$Y_{ijkl} = \mu + B_i + M_j + F_k + MF_{jk} + C(MF)_{l(jk)} + e_{ijkl}$$

where, Y_{ijkl} is the observed value, μ is the overall mean, B_i is the effect of block i, M_j is the effect of the male parent j, F_k is the effect of the female parent k, MF_{jk} is the interaction effect between male and female parent jk (i.e. family), $C(MF)_{l(jk)}$ is the effect of clone l within family, and e_{ijkl} is the random error. All effects except μ were considered random.

The general combining ability (GCA) or the effects of parents was extracted from this model to provide the relative phenotypic contributions of the male parent (M_j) and female parent (F_k) to family progeny. The deviation of progeny trait values relative to their parents (midparent heterosis, MPH) was calculated for each family as $[(F_1 - MPV) / MPV] \times 100$. Family 425 was removed from MPH calculations because the male parent was not present.

To evaluate the total variance explained by each predictor variable in multiple linear regression analyses, Lindeman, Merenda, and Gold (LMG) (Lindeman, et al. 1980), First and Last, Genizi (Genizi, 1993), and Pratt (Pratt, 1987) metrics were used to order predictors and decompose R^2 in the relaimpo (Relative Importance for Linear Regression) package (Grömping, 2006). In order to assess which regressors are different in terms of relative importance, confidence intervals were calculated using 1,000 bootstrapped replicates (Bonferroni CI=95%). Pratt and LMG indices were used to partition the additive properties of R^2, calculated as the sum of their individual importance irrespective of the correlation among predictor variables.

2.4 Results

2.4.1 Family Variation in REC and Biomass Production

The means of REC significantly differed by up to 34% among the eight full-sib families (Figure 2.1, Table 2.3). The triploid family 430 (*S. miyabeana* × *S. viminalis*) displayed the greatest REC means (182.8 nF), while the triploid family 415 (*S. purpurea* × *S. miyabeana*) has the lowest REC with a mean of 120.8 nF (Figure 2.1, Table 2.3). The effect of progeny within family accounted for a majority of the variance of REC, and the effect of the female parent was almost twice that of the effect of the male parent. Midparent heterosis for REC was observed in three of seven families (421, 423, 430), but not in families 082 (F_1 *S. purpurea* × *S. purpurea*), 317 (F_2 *S. purpurea* × *S. purpurea*), 415 (*S. purpurea* × *S. miyabeana*) and 407 (*S. purpurea* × *S. viminalis*) (Figure 1.1). Since one of the parents of family 425 was not included, heterosis was not calculated for that family.

Combinations of parents with high GCA values should theoretically produce progeny with high absolute trait values. The female and male *S. viminalis* (07-MBG-5027 and 'Jorr') and the female *S. miyabeana* (01-200-006) parent clones had the greatest REC GCA of family parents (Table 2.4). In absolute terms, the *S. purpurea* female clone 94006 had the greatest GCA of all parents. Midparent heterosis for REC was observed to be the most dramatic for interspecific crosses of these parents, i.e. families 430 (t=4.09, P<0.001) (01-200-006 × 'Jorr'), 423 (t=3.56, P=0.003) (07-MBG-5027 × 01-200-003), and 421 (t=2.59, P=0.02) (07-MBG-5027 × 94001).

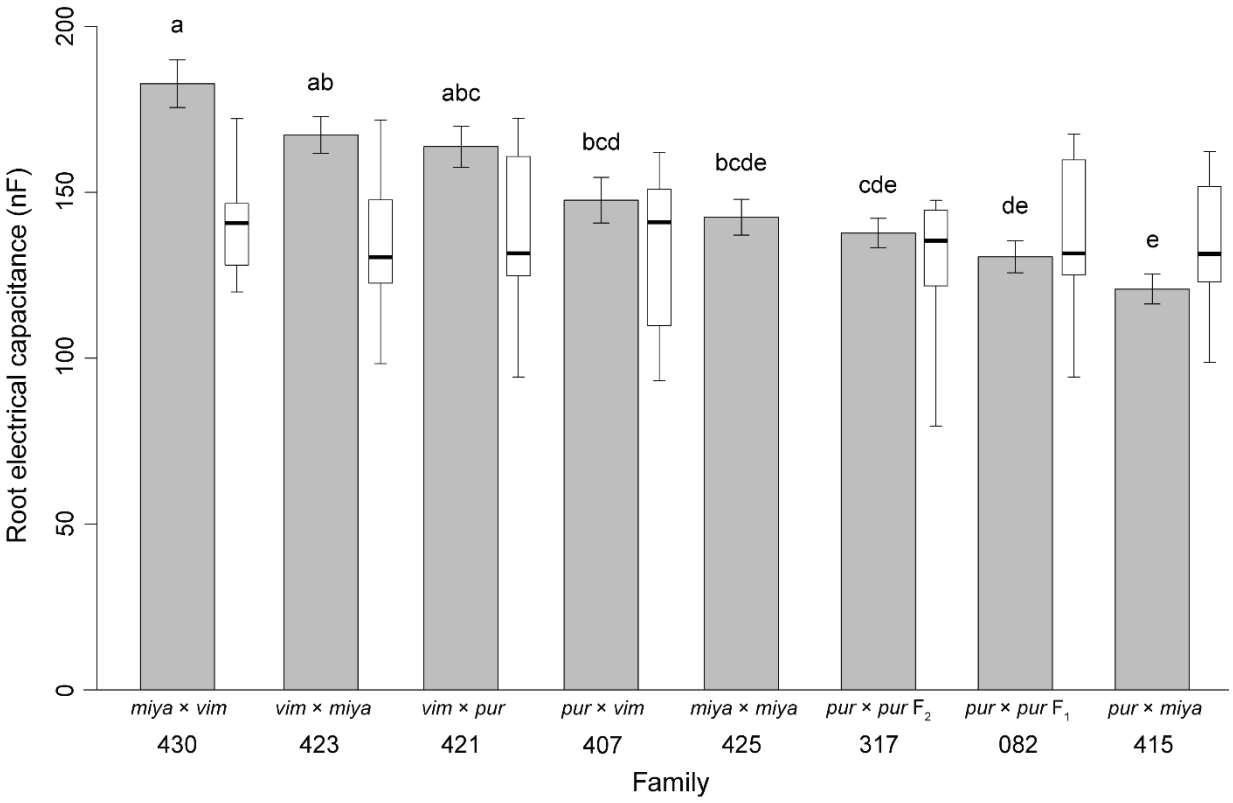

Figure 2.1 Root electrical capacitance (REC) family means and errors (±SE) of four replicates as well as their respective Tukey's honest significant difference (HSD) at a 95% level of confidence. Outlined boxplots to the right of bars represent the corresponding midparent value (±SD) for the parents of each family. Midparent means within boxplots are denoted by bold horizontal lines. The male parent of the 425 family was not present in this trial, thus the midparent boxplot of family 425 is not shown. The species pedigree of each family is abbreviated above family identifiers, where '*pur*' = *Salix purpurea*, '*vim*' = *S. viminalis*, and '*miya*' = *S. miyabeana*. The direction of the cross is designated as female parent × male parent species.

Table 2.3 Root biomass electrical capacitance (REC nF) family means, standard errors (± SE), and Tukey honest significant difference (HSD) at a 95% level of confidence. A *t*-Test of significance was carried out on family progeny and the respective midparent value to ascertain the magnitude of midparent heterosis (MPH), where the hypothesized differences in means is zero. Families 430, 424, and 421 display midparent heterosis for REC, whereas all other families are not significantly different ($P < 0.05$) from their midparent values. The male parent of the 425 family was not present in this trial, thus midparent heterosis of family 425 is not shown.

Family	**Mean**	**± SE**	**HSD**	**MPH**	***t-value***	***p-value***
430	182.8	7.2	a	36.3	4.09	<0.001
423	167.3	5.6	ab	25.1	3.56	0.003
421	163.8	6.2	abc	20.5	2.59	0.020
407	147.7	6.9	bcd	7.4	0.84	0.415
425	142.5	5.4	bcde	–	–	–
317	137.7	4.5	cde	6.9	0.99	0.337
082	130.6	4.8	de	-4.2	-0.59	0.565
415	120.8	4.5	e	-9.9	-1.56	0.142

Table 2.4 Family parent general combining abilities (GCA) of root electrical capacitance (REC), stem dry weight biomass (SDW), and leaf dry weight biomass (LDW).

	REC	SDW	LDW
Females			
01-200-006	5.84	-1.26	3.58
07-MBG-5027	18.62	4.72	3.84
94006	-17.95	-3.78	-3.43
'Wolcott'	-6.51	0.32	-3.99
Males			
01-200-003	-4.05	0.01	1.40
SX64	-7.80	-3.25	-0.49
'Jorr'	14.40	2.96	0.16
94001	-0.48	0.10	-0.63
'Fish Creek'	-3.02	0.18	-0.44

2.4.2 Correlation of Biomass Traits with REC

Initial estimates of the predictive accuracy of RDW using REC (adjusted $R^2 = 0.77$, $P = 0.001$) were cross-validated using *K*-fold ($k = 5$) and leave-one out (LOOCV) (N-1) methods ($R^2 = 0.71$, RMSE = 5.1) (Figure 2.2). For all traits analyzed, significant REC correlation coefficients ($P<0.05$) ranged from 0.11 (SPAD) to 0.88 (RDW). The REC trait was positively correlated with many above- and belowground biomass traits (Table 2.2). Aboveground traits that most strongly correlate with REC were SDW, LDW, SSL, and MSA (Table 2.5). Belowground traits that most strongly correlated with REC were RDW, CDW, and CDIA (Table 2.5).

2.4.3 Model Selection and Predictor Comparison

Correlations of above- and belowground biomass-related traits with REC and SDW were performed in order to rank predictors in multiple regression models. To avoid any potential issues with collinearity in downstream analysis, repeated trait measurements were reduced to a single representative measurement based on the proportion of model variance it explained. For instance, the third of five phenology measurements (12 DAP) and the sixth of nine stem length and stem number measurements (48 DAP) were chosen because they had the strongest correlation with REC and/or SDW.

Important predictors of RDW were selected from a total of 15 cutting and root traits from individuals selected along the distribution of REC readings using stepwise regression (Step AIC, direction=both) of the linear model fit. Owing to the significant contribution of REC and SRN predictor variables, the reduced model, RDW ~ REC + CDIA + PRN + SRN, explains a substantial portion of the variation in RDW (multiple $R^2 = 0.85$, $P < 0.001$) (Table 2.6).

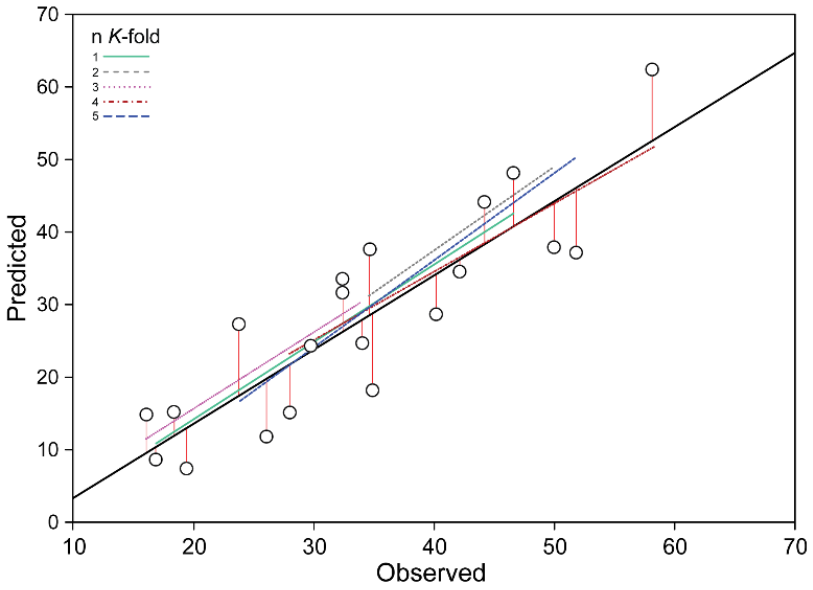

Figure 2.2 Linear regression and *K*-fold cross-validation scatter plot of observed root dry weight biomass (RDW) predicted by root electrical capacitance (REC). The linear regression is $y = 0.21x - 6.91$ (adjusted $R^2 = 0.77$). Vertical red lines perpendicular to the regression line (solid black) represent the residual of the model fit and colored lines represent *K*-fold sampling.

Table 2.5 Pearson correlation coefficients of above and belowground traits with root electrical capacitance (REC) and root dry weight (RDW).

	REC			RDW		
Trait	***r***	**CI**	***P***	***r***	**CI**	***P***
Above ground						
SDW	0.72	0.66 – 0.75	<0.001	0.52	0.29 – 0.75	0.019
LDW	0.56	0.49 – 0.62	<0.001	0.78	0.51 – 0.91	<0.001
SSL	0.53	0.45 – 0.59	<0.001	0.59	0.22 – 0.82	0.005
MSL	0.37	0.28 – 0.45	<0.001	0.69	0.35 – 0.86	<0.001
MXL	0.44	0.35 – 0.31	<0.001	0.78	0.52 – 0.91	<0.001
RGR	0.42	0.33 – 0.49	<0.001	0.33	0.27 – 0.44	0.005
MSA	0.23	0.14 – 0.32	<0.001	0.31	–	0.178
SSA	0.39	0.31 – 0.46	<0.001	0.20	–	0.406
LFA	0.09	–	0.092	0.32	–	0.159
PSN	0.22	0.13 – 0.31	<0.001	0.23	–	0.310
ASN	0.19	0.10 – 0.28	<0.001	0.64	0.29 – 0.85	0.002
SPD	-0.11	0.01 – 0.20	<0.001	-0.34	–	0.139
PHE	0.38	0.29 – 0.46	<0.001	0.58	0.18 – 0.81	0.007
MSD	0.13	0.03 – 0.22	<0.001	0.33	–	0.148
SSD	0.39	0.31 – 0.47	0.0091	0.19	–	0.414
SLA	-0.23	0.13 – 0.31	<0.001	-0.56	0.15 – 0.80	0.010
LMA	0.24	0.20 – 0.30	<0.001	0.59	0.13 – 0.79	0.007
Below ground						
REC				0.88	0.71 – 0.95	<0.001
RDW	0.88	0.71 – 0.95	<0.001			
CDIA	0.65	0.30 – 0.85	0.001	0.67	0.33 – 0.86	0.001
CDW	0.67	0.33 – 0.86	<0.001	0.70	0.37 – 0.84	<0.001
RDIA	0.20	0.27 – 0.59	0.002	0.05	–	0.982
PRN	0.64	0.29 – 0.84	0.016	0.50	0.08 – 0.77	0.023
TRN	0.59	0.20 – 0.81	0.041	0.53	0.12 – 0.79	0.005
SRN	0.34	0.12 – 0.67	0.006	0.45	0.03 – 0.74	0.041

Table 2.6 Model coefficients of important predictors selected from a total of 15 cutting and root traits for use in multiple regression for root dry weight (RDW) using stepwise AIC (multiple R^2=0.85, adj R^2=0.81).

Trait	SS	Residual SS	AIC	*F*
REC	1024.38	1611.00	95.778	26.193***
CDI	70.00	656.630	77.828	1.790
PRN	111.97	698.600	79.067	2.863
SRN	191.29	777.910	81.218	4.891*

*** Significant at p<0.001
** Significant at p<0.01
* Significant at p<0.05

Variable selection for SDW was conducted using common relative importance indexing methods. Confidence intervals for each model were produced using 1,000 bootstrap replicates at a 95% level of significance. The LMG method is the most commonly-used metric in multiple regression. With regards to ranking predictors based on their relative importance, Genizi and Pratt CIs were equivalent to the LMG method (Table 2.7). Outcomes from the Genizi predictor ranking method were excluded as both predictor importance and contributions were nearly identical to the LMG method. The Pratt index was selected to order predictor relative importance in multiple regression analyses as it explains the total variance as the sum of predictor R^2.

Relative importance metrics for predictor variables did not vary greatly among model orderings (i.e. rank), whereas in contrast to within-trait R^2 and bootstrapped CIs (boot=1000, Bonferroni CI = 95%), the magnitude of the difference in variation explained in the predictors did vary significantly (Figure 2.3). The proportion of variance that could be explained using the seven predictors in the final allometric model, SDW ~ REC + SSL + MSL + SSD + ASN + PSN + SPD, was 68.5% (Table 2.7). A large proportion of the variance is explained by the relative contribution of REC (R^2 = 0.42), SSL (R^2 = 0.14) and MSL (R^2 = 0.15).

Table 2.7 Relative importance metrics for multiple linear regression of 10 traits important for biomass production to stem dry weight biomass (SDW). Each model describes the variance explained by predictor variables as a percentage of R^2 (summed to 1). Relative contributions of individual predictors in each model are ranked based on confidence intervals (CI) produced using 1000 bootstrap replicates at a significance level of 95%.

	LMG			First			Last			Pratt		
Trait	**R^2**	**CI**	**Rank**	**R^2**	**CI**	**Rank**	**R^2**	**CI**	**Rank**	**R^2**	**CI**	**Rank**
REC	0.41	0.37 – 0.48	a	0.41	0.36 – 0.47	a	0.50	0.31 – 0.65	a	0.42	0.35 – 0.49	a
MSL	0.15	0.10 – 0.20	bcd	0.14	0.09 – 0.19	bcd	0.07	0.04 – 0.18	bcdefg	0.15	0.10 – 0.21	bcd
SSL	0.15	0.11 – 0.19	bcd	0.18	0.14 – 0.22	bc	0.10	0.05 – 0.18	bcdefg	0.14	0.11 – 0.18	bcd
SSD	0.12	0.09 – 0.16	bcde	0.15	0.11 – 0.19	bcd	0.07	0.03 – 0.14	bcdefg	0.12	0.09 – 0.15	bcde
ASN	0.07	0.04 – 0.11	defg	0.05	0.02 – 0.10	efg	0.09	0.04 – 0.17	bcdefg	0.07	0.04 – 0.11	defg
PSN	0.06	0.05 – 0.09	efg	0.04	0.02 – 0.08	efg	0.06	0.02 – 0.14	cdefg	0.06	0.05 – 0.08	efg
SPD	0.04	0.02 – 0.07	efg	0.02	0.01 – 0.05	fg	0.01	0.05 – 0.19	bcdefg	0.04	0.02 – 0.07	efg

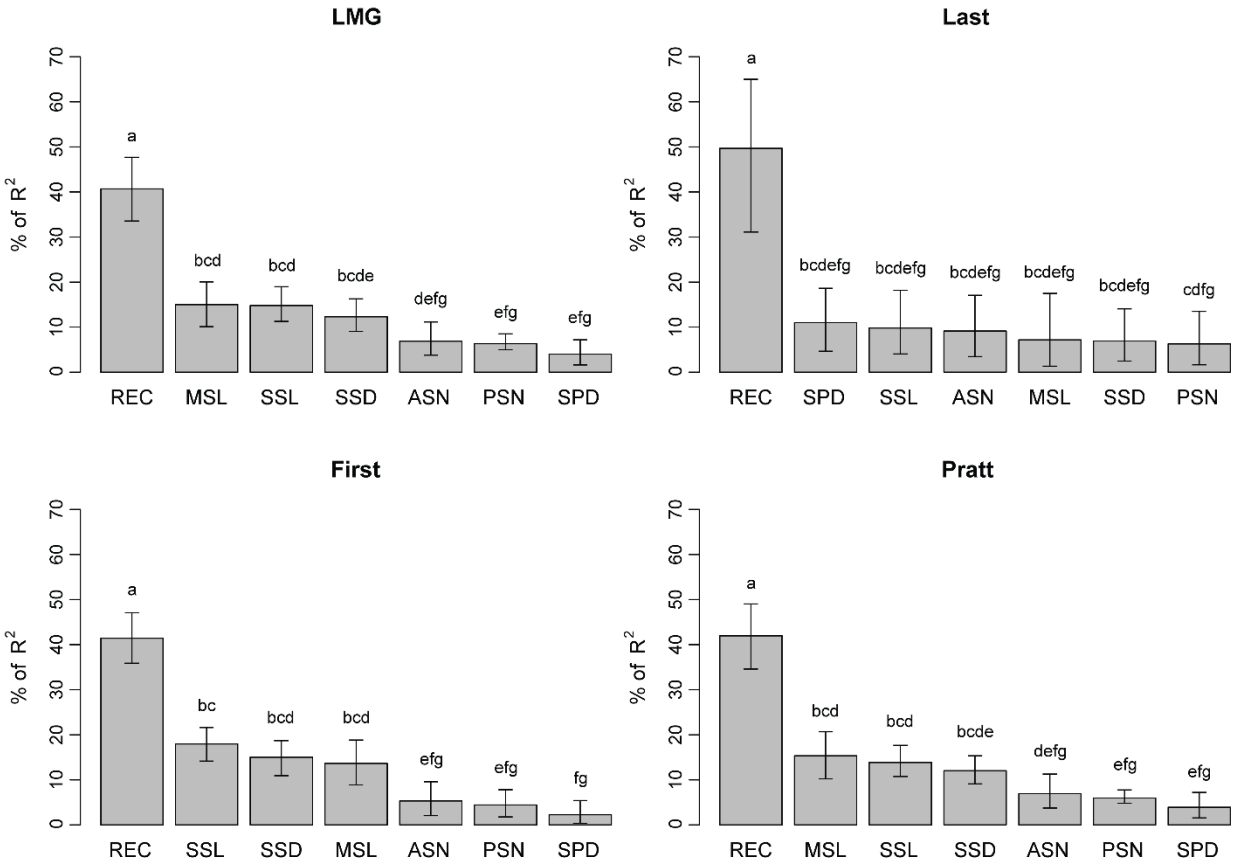

Figure 2.3 Relative importance predictor index metrics for total stem dry weight biomass (SDW). Barplots of each model index describes the variance explained as a percentage R^2 (summed to 100%). For each predictor, confidence intervals (CI) were produced using 1000 bootstrap replicates at a 95% level of significance. Low percentages of total R^2 or CIs approaching zero describe unimportant variable contributions (e.g., SPD, "First" index method).

When REC is not included as a predictor variable in multiple regression, only 55.9% of the variation in SDW could be explained by the remaining variables; a difference of 12.6%. In each model index, REC was found to be significantly different compared to all other predictor variables and explained the greatest proportion of variance in each model index.

2.5 Discussion

There are a number of key traits that could be selected for in shrub willow breeding programs if a rapid method for phenotyping root traits was available, including rate of rooting

establishment, ability to compete with weeds, drought tolerance, and total carbon sequestration ability. The partitioning of photoassimilate from foliage to stem and root biomass is largely dependent upon a complex network of photosynthetic and transport processes, as well as other morphological, architectural, and environmental factors. It is naïve to assume that the root:shoot ratio is constant among segregating progeny, rather it is likely there are genotypic differences in the accumulation of root biomass in proportion to aboveground biomass. For the reason that aboveground biomass was harvested prior to taking REC measurements in this study, any confounding factors that may have been introduced into analyses were avoided. While it is true that the quantitative nature of REC does not offer direct insight into root morphology or architecture, the specific morphologically-based variances in REC need not be known if it is still a good predictor of RDW.

Increased vigor can be observed in intraspecific crosses of shrub willow but the effect is generally more dramatic in the progeny of interspecific crosses (Serapiglia et al., 2014), especially in triploid progeny derived from the hybridization of tetraploid and diploid parents (Serapiglia et al., 2015; Fabio et al. 2016). Likewise, we show that the interspecific families 421, 423, and 430 had significantly higher means for REC compared to their respective midparent values (Figure 2.1; Table 2.3). The interspecific diploid family 421 was the result of a wide cross between parent species *S. viminalis* (Section *Vetrix*) and *S. purpurea* (Section *Helix*). The interspecific triploid families 430 and 423 were derived from parents of contrasting ploidy-level as well as Section membership within *Vetrix*, which may explain higher-levels of midparent heterosis for this trait within these families.

We find that REC is highly correlated with RDW among all progeny and parent genotypes in this study; comparable to previous correlations found for one cultivar of shrub

willow (*S. viminalis* × *S. schweinii* 'Olof') (Pitre *et al*., 2010) and hybrid poplar (*Populus deltoidies* × *P. nigra*) (Preston et al., 2004). Working with four commercial shrub willow cultivars, Cunniff et al. (2015) demonstrated LFA to have the strongest correlation with RDW (R^2=0.49) compared with other aboveground traits in a field setting. In this study, LFA was weakly correlated with RDW and only significantly correlated with REC (R^2 = 0.40, P = 0.005) in the diploid family 407 (*S. purpurea* × *S. viminalis*). Besides the strong correlation of RDW with REC, RDW had a strong linear and positive correlation with LDW and LMA, and a strong negative correlation to SLA (Table 2.5). Although LDW was strongly correlated with SDW (R^2 = 0.81, P < 0.001), it was excluded as a predictor variable in multiple regression analyses because it is a destructive measurement and not manageable to collect leaf biomass in a field setting.

Estimation of root biomass relative growth rates may be modeled from multi-temporal data with minimal effort. The root system of shrub willow typically resides in the upper 20-30 cm of the soil and fine roots in the upper 10 cm (Pacaldo et al., 2013), which may find worthy implementation of REC as a long-term indicator of root development in research field trials. In addition, there is also potential to investigate the relationship of REC with the occurrence and distribution of ectomycorrhizal and arbuscular mycorrhizal populations among short rotation plantations and natural stands of shrub willow (Dhillion 1994; Cseresnyés et al., 2013).

Measurement of woody plant root biomass is an extremely laborious and expensive task as it requires careful washing, sieving, storing, drying, and weighing root tissues. The REC method provides an efficient, reproducible, and non-destructive alternative to more traditional 'shovel-omics' techniques. Still, the efficiency of REC must be improved and the equipment itself scaled-up in order to withstand heterogeneous soil profiles and thick stools of mature plants

in the field. For instance, Ellis et al. (2013) utilized the electrical capacitance 4T method as an indicator of root length in *Pinus, Eucalyptus,* and *Corymbia* forest tree plantations, resolving capacitance somewhat related to root mass within species but a poor indicator of diameter at breast height, root mass, or root length across species, age, or sites. Technological improvements in field-based root phenotyping (Rautenbach et al., 2013; Meister et al., 2014) will hopefully allow for high-throughput quantification of root architecture and biomass at a reasonable efficiency and cost.

Compared to annual crops, it is likely that the electrical capacitances of perennial shrub willows will change over time as a result of root suberization, repeated coppicing, and seasonal variation of biotic and abiotic stressors. As implied, future work must be accomplished in order to determine if the strong relationship between REC and biomass-related traits on second and third-year growth are equally as strong as in the first-year. It is likely that over multiple coppice cycles, environment will play an important role in determining the genotypic stability of this trait in shrub willow. Nevertheless, the strength of REC prediction estimates of above- and belowground dry biomass is sufficient to conclude it is not only valuable but is adaptable across *Salix* species, ploidy-level, and hybrid pedigree.

2.6 REFERENCES

Alfons A. 2012. cvTools: Cross-validation tools for regression models. Comprehensive R Archive Network. Available at: https://CRAN.R-project.org/package=cvTools

Argus GW. 2005. Guide to the identification of the Genus *Salix* (willow) in New England and New York. Delta Institute of Natural History, Bowdoin, ME.

Argus GW. 2012. *Salix* (willows) in the New World: a guide to the interactive identification of native and naturalized taxa using Intkey (DELTA). Prairie and Northern Plant Diversity Centre, Devonian Botanical Garden, Edmonton, Alberta.

Bates D, Maechler M, Bolker B, Walker S. 2015. lme4: Mixed-effects modeling with R. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67: 1–48.

Bonosi L, Ghelardini L, Weih M. 2013. Towards making willows potential bio-resources in the South: Northern *Salix* hybrids can cope with warm and dry climate when irrigated. *Biomass and Bioenergy*, 51: 136–144.

Cao Y, Repo T, Silvennoinen R, Lehto T, Pelkonen P. 2011. Analysis of the willow root system by electrical impedance spectroscopy. *Journal of Experimental Botany*, 62: 351–358.

Cseresnyés I, Takács T, Végh KR, Anton A, Rajkai K. 2013. Electrical impedance and capacitance method: A new approach for detection of functional aspects of arbuscular mycorrhizal colonization in maize. *European Journal of Soil Biology*, 54: 25–31.

Cunniff J, Purdy SJ, Barraclough TJP, Castle M, Maddison AL, Jones LE, Shield IF, Gregory IS, Karp A. 2015. High yielding biomass genotypes of willow (*Salix* spp.) show differences in below ground biomass allocation. *Biomass and Bioenergy*, 80: 114–127.

Currie PO, Hilken TO, White RS. 1987. Evaluation of a single probe capacitance meter for estimating herbage yield. *Society for Range Management,* 40: 537–541.

Dalton FN. 1995. In-situ root extent measurements by electrical capacitance methods. *Plant and Soil,* 173: 157–165.

Dhillion SS. 1994. Ectomycorrhizae, arbuscular mycorrhizae and *Rhizoctonia* sp. of alpine and boreal *Salix* spp. in Norway. *Arctic and Alpine Research,* 26: 304–307.

Dietrich RC, Bengough AG, Jones HG, White PJ. 2012. A new physical interpretation of plant root capacitance. *Journal of Experimental Botany,* 63: 6149–6159.

Dietrich RC, Bengough AG, Jones HG, White PJ. 2013. Can root electrical capacitance be used to predict root mass in soil? *Annals of Botany,* 112: 457–464.

Ellis TW, Murray W, Paul K, Kavalieris L, Brophy J, Williams C, Maass M. 2013. Electrical capacitance as a rapid and non-invasive indicator of root length. *Tree Physiology,* 33: 3–17.

Fabio ES, Volk TA, Miller RO, Serapiglia MJ, Gauch HG, Van Rees KCJ, Hangs RD, Amichev BY, Kuzovkina YA, Labrecque M, Johnson GA, Ewy RG, Kling GJ, Smart LB. 2016. Genotype × environment interaction analysis of North American shrub willow yield trials confirms superior performance of triploid hybrids. *GCB Bioenergy*, doi:10.1111/gcbb.12344

Genizi A. 1993. Decomposition of R^2 in multiple regression with correlated regressors. *Statistica Sinica* 3: 407–420.

Grömping U. 2006. Relative importance for linear regression in R: the package relaimpo. *Journal of Statistical Software,* 17: 1–27.

Karp A, Hanley SJ, Trybush SO, Macalpine W, Pei M, Shield I. 2011. Genetic improvement of willow for bioenergy and biofuels. *Journal of Integrative Plant Biology,* 53: 151–165.

Kormanek M, Tomasz G. 2015. Modification of the tree root electrical capacitance method under laboratory conditions. *Tree Physiology,* 36: 121–127.

Kuzovkina YA, Romero MA, Charles J, Hurst J, Mcivor I, Karp A, Trybush S, Labrecque M, Teodorescu TI, Singh NB, Smart LB, Volk TA. 2008. *Salix*: botany and global horticulture. *Horticultural Reviews,* 34: 447–489.

Lauron-Moreau A, Pitre FE, Argus GW, Labrecque M, Brouillet L. 2015. Phylogenetic relationships of American willows (*Salix* L., Salicaceae). *PLoS ONE,* 10: e0138963.

Lekas TM, Macdougall GR, Maclean DA, Thompson RG. 1990. Seasonal trends and effects of temperature and rainfall on stem electrical capacitance of spruce and fir trees. *Canadian Journal of Forestry,* 20: 970–977.

Lindeman RH, Merenda PF, Gold RZ. 1980. Introduction to bivariate and multivariate analysis, 119. Longman Higher Education, Glenview, IL.

Meister R, Rajani MS, Ruzicka D, Schachtman DP. 2014. Challenges of modifying root traits in crops for agriculture. *Trends in Plant Science,* 19: 779–788.

Pacaldo RS, Volk TA, Briggs RD. 2013. No significant differences in soil organic carbon contents along a chronosequence of shrub willow biomass crop fields. *Biomass and Bioenergy,* 58: 136–142.

Pitre FE, Brereton NJB, Audoire S, Richter GM, Shield I, Karp A. 2010. Estimating root biomass in *Salix viminalis* × *Salix schwerinii* cultivar “Olof” using the electrical capacitance method. *Plant Biosystems,* 144: 479–483.

Pratt JW. 1987. Dividing the indivisible: using simple symmetry to partition variance explained. In: Pukkila T, Puntanen S, editors, Proceedings of the Second International Conference in Statistics, 245–260. University of Tampere, Tampere, Finland.

Preston GM, Mcbride RA, Bryan J, Candido M. 2004. Estimating root mass in young hybrid poplar trees using the electrical capacitance method. *Agroforestry Systems*, 60: 305–309.

R Development Core Team. 2011. R: a language and environment for statistical computing. R foundation for statistical computing, 1–409. Vienna, Austria. Available at: https://www.R-project.org/.

Rajkai K, Végh KR, Nacsa T. 2005. Electrical capacitance of roots in relation to plant electrodes, measuring frequency and root media. *Acta Agronomica Hungarica*, 53: 197–210.

Rautenbach C, Mudde RF, Yang X, Melaaen MC, Halvoresen BM. 2013. A comparative study between electrical capacitance tomography and time-resolved X-ray tomography. *Flow Measurement and Instrumentation*, 30: 34–44.

Repo T, Laukkanen J, Silvennoinen R. 2005. Measurement of the tree root growth using electrical impedance spectroscopy. *Silva Fennica*, 39: 159–166.

Serapiglia MJ, Gouker FE, Smart LB. 2014. Early selection of novel triploid hybrids of shrub willow with improved biomass yield relative to diploids. *BMC Plant Biology*, 14, doi: 10.1186/1471-2229-14-74.

Serapiglia MJ, Gouker FE, Hart JF, Unda F, Mansfield SD, Stipanovic AJ, Smart LB. 2015. Ploidy level affects important biomass traits of novel shrub willow (*Salix*) hybrids, *Bioenergy Research*, 8: 259–269.

Smart LB, Cameron KD. 2012. Shrub willow. In: C. Kole, S. Joshi S., and D. Shonnard (Eds.), Handbook of bioenergy crop plants, 687–708. Taylor and Francis Group, Boca Raton, FL.

Walther GR, Post E, Convey P, Menzel A, Parmesean C, Beebee TJC, Fromentin JM, Hoegh-Guldberg O, Bairlein F. 2002. Ecological responses to recent climate change. *Nature,* 416: 389–95.

Whiteman G, Hope C, Wadhams P. 2013. Vast costs of Arctic change. *Nature* 499: 401–403.

Whitlow TH, Bassuk NL, Ranney TG, Reichert DL. 1992. An improved method for using electrolyte leakage to assess membrane competence in plant tissues. *Plant Physiology,* 98: 198–205.

CHAPTER 3

HETEROSIS FOR BIOMASS-RELATED TRAITS IN NOVEL TRIPLOID HYBRIDS OF SHRUB WILLOW (*SALIX* SPP.)

3.1 Abstract

Hybridization is key in the improvement of shrub willow (*Salix* spp.) bioenergy crops, because hybrids often display heterosis for yield. Development of high-yielding genotypes will require numerous rounds of evaluation, selection, and hybridization. A primary breeding objective for the improvement of shrub willow as a bioenergy crop is to develop full-sib populations and evaluate entire families in family-based selection trials. Improving the efficiency of evaluation and selection must be considered to optimize future efforts. The extent to which intra- and interspecific F_1 and F_2 shrub willow (*Salix* spp.) exhibit heterosis for biomass yield was examined utilizing a suite of biomass, foliar, and physiological traits collected over the course of 12 weeks in the greenhouse and over two years in the field. Triploid families generated from diploid *S. viminalis* and tetraploid *S. miyabeana* showed the highest levels of heterosis for harvestable biomass and biomass-related growth traits in the greenhouse and in the field. While intraspecific *S. purpurea* diploids showed low levels of heterosis for these traits, interspecific diploids did exhibit moderate levels of heterosis in greenhouse experiments. Differences between growth trials can largely be explained by pest incidence, for which interspecific diploids were negatively impacted. Beyond biomass yield, family-level differences in heterosis for stem growth, foliar, and physiological traits are discussed.

3.2 Introduction

Shrub willow (*Salix* spp., Salicaceae) are vigorous woody perennials bred as feedstocks for dedicated biofuel production (Smart and Cameron, 2012). Found within marginal and riparian habitats, the range of *Salix* extends from the arctic plains to the subtropics, and more than 350 species have been described (Lauron-Moreau et al., 2015, Argus, 1997). *Salix* are dioecious outcrossing species that are suspected to be both entomophilous and anemophilous (Argus, 1974), but chiefly rely on the former (Tamura and Kudo, 2000). As *Salix* is particularly amenable to wide-hybridization, taxonomic characterization within the genus has been an enduring challenge for botanists and plant breeders alike (Percy et al., 2014). In sympatric populations of *Salix*, members within the same section will often hybridize, which can generate mixed populations of both pure species and species hybrids (Hardig et al., 2000). In addition, vegetative clonal propagation can be a significant contributor to population structure in *Salix*, and has been shown to be fairly common in naturalized stands of North American *S. purpurea* (Lin et al., 2009). Beyond the tremendous ecological amplitude, the heterogeneity and adaptive plasticity of *Salix* delivers a prodigious source of germplasm for genetic improvement.

The domestication of *Salix* traces back to the Swedish geneticist Nils Heribert-Nilsson's early cytological studies of *S. viminalis* × *S. caprea* hybrids in the 1920's (Heribert-Nilsson, 1918). Shortly thereafter, willow conservation and breeding was principally led by H.P. Hutchinson and K.G. Scott for nearly 30 years at the Long Ashton Experiment Station in the UK. Since the 1970's, breeders have maintained a goal of producing fast-growing shrub willow bioenergy feedstock cultivars that are high-yielding, genetically diverse, resistant to pests and diseases, and amenable to marginal sites, without competing with food crops (Karp et al., 2011). A thorough review of willow botany and breeding can be found in (Kuzovkina et al., 2008).

While most complex traits are generally considered to be additively inherited, deviations from the midparent value in the F_1 can result in either hybrid vigor or hybrid necrosis (Birchler, 2006). Since the Green Revolution, the phenomenon of hybrid vigor has been exploited in crop systems, more than doubling global commodity yields in only a few decades. In order to advance the adaptive capacity of US agricultural and energy sectors in response to climate change, plant breeders must develop regionally dedicated and sustainable bioenergy crops displaying hybrid vigor. Shrub willow has great potential as a competitive bioenergy feedstock that can directly substitute for fossil fuels with great potential for yield increases through species hybridization.

Hybridization is a key component in the development of shrub willow bioenergy crops, as hybrids often display heterosis for yield (Zsuffa et al., 1984). By careful selection, hybridization, and phenotypic evaluations, substantial improvements have been made in shrub willow biomass yield (Volk et al., 2011) and quality (Fabio et al., 2017b). While heterosis has been realized in intraspecific crosses (Cameron et al., 2008), it is more pronounced in triploids derived from the hybridization of diploid and tetraploid parents (Fabio et al., 2017a). Previous shrub willow yield trials have shown that elite triploid hybrid cultivars produce higher biomass yield compared to diploids and exceed or are not significantly different from tetraploids (Fabio et al., 2016). In addition, there is evidence that triploids have the potential to produce more cellulose per unit area because of higher yields and/or cellulose content (Fabio et al., 2017b). Nevertheless, little is known about the genetic mechanisms responsible for the phenomenon of heterosis in triploid shrub willow (Serapiglia et al., 2014).

The inherent time and investment associated with cultivar development in woody perennial feedstocks require efficient population development and phenotypic evaluation methods to significantly increase gains in biomass yield. Evaluation of large populations of

progeny will be necessary to make significant gains, which requires early screening to reduce the costs associated with long-term field trials. The families referenced in this study were generated from crosses made between both ploidy and Section within the genus *Salix*, involving diploid *S. purpurea* and tetraploid *S. miyabeana* in the Section *Helix* Dumont, and diploid *S. viminalis* in the Section *Vimen* (formerly *Viminella* Seringe). The main objectives of this study were (1) to compare and contrast extensive phenotypic data collected among eight intra- and interspecific shrub willow families, (2) to determine the extent to which these families display heterosis for biomass-related traits and biomass yield, and (3) to develop family-specific models for yield using multivariate techniques.

3.3 Materials and Methods

3.3.1 Population Development

A total of eight full-sib F_1 and F_2 families were generated from crosses between diploid and tetraploid parents representing three *Salix* species: *S. purpurea* Sect. Helix (2n=2x=38), *S. viminalis* Sect. Vimen (2n=2x=38), and *S. miyabeana* Sect. Helix (2n=2x=76). The full-sib F_1 *S. purpurea* family 82 was generated from a cross between female 94006 and male 94001, both collected from naturalized *S. purpurea* populations in upstate NY. Two F_1 offspring from this cross, female *S. purpurea* 'Wolcott' (9882-41) and male *S. purpurea* 'Fish Creek' (9882-34), were crossed to generate the full-sib F_2 *S. purpurea* family 317. The female *S. purpurea* genotype 94006 was crossed with the male *S. viminalis* 'Jorr' to generate the interspecific diploid family 407, and a cross between *S. viminalis* 07-MBG-5027 and the male *S. purpurea* genotype 94001 generated the 'pseudo-reciprocal' interspecific family 421. Female diploid genotypes 94006 and 07-MBG-5027 were separately crossed with tetraploid *S. miyabeana* male 01-200-003, to generate the interspecific triploid families 415 and 423, respectively. Triploid family 430

was generated from a cross between the tetraploid *S. miyabeana* female 01-200-006 and diploid *S. viminalis* male 'Jorr'. Finally, the intraspecific *S. miyabeana* tetraploid family 425 resulted from a cross between female 01-200-006 and male 'SX64'. All family progeny individuals and their parents were planted in nursery beds at Cornell AgriTech, Geneva, NY.

Table 3.1 Description of intra- and interspecific shrub willow family parents, their pedigree, generation, and ploidy-level.

Family	Female	Male	Pedigree	Ploidy	Total	Final
82	94006	94001	*S. purpurea* × *S. purpurea*	2*x*	423	104
317	Wolcott	Fish Creek	*S. purpurea* × *S. purpurea*	2*x*	493	482
407	94006	Jorr	*S. purpurea* × *S. viminalis*	2*x*	282	100
421	07-MBG-5027	94001	*S. viminalis* × *S. purpurea*	2*x*	244	100
415	94006	01-200-003	*S. purpurea* × *S. miyabeana*	3*x*	306	100
423	07-MBG-5027	01-200-003	*S. viminalis* × *S. miyabeana*	3*x*	75	63
430	01-200-006	Jorr	*S. miyabeana* × *S. viminalis*	3*x*	39	23
425	01-200-006	SX64	*S. miyabeana* × *S. miyabeana*	4*x*	136	100

3.3.2 Greenhouse Design

Parent genotypes and randomly chosen progeny from the eight families described above were grown from stem cuttings (20 cm) in 12-L plastic pots with peat moss-based potting mix (Fafard, Agawam, MA) to evaluate growth traits under greenhouse conditions over the course of 12 weeks. Families consisted of 12 progeny individuals and their parents, for a total of 104 genotypes (Table 3.1). One exception is that the male parent SX64 of the intraspecific *S. miyabeana* family 425 was not included. Plot was defined as a single cutting planted in a pot, which were arranged in a randomized complete block design with four replicate blocks. Two blocks were located on benches in one greenhouse with the other two blocks in an adjacent greenhouse set for identical growing conditions. Supplemental greenhouse lighting was provided on a 14-h day : 10-h night regimen with max daytime temperature of 26°C and a nighttime

temperature of 18°C. Liquid fertilizer (Peter's 15-16-17 Peat-Lite Special®, Scott's, Marysville, OH) was applied weekly after week four according to manufacturer recommendations.

3.3.3 Field Design

The field trial was established at Cornell AgriTech in Geneva, NY in a randomized complete block design with four replicate blocks of three-plant plots. To avoid edge effects, *S. purpurea* genotypes 'Fish Creek' and 94006 were planted as border rows along the east and west sides of the trial, respectively, and the north and south ends were buffered by a single row of genotype 94006. Within-row spacing was 0.4 m and spacing between rows was 1.82 m. The soil at the field site is Odessa silt loam with a depth to water table of 25 to 45 cm. For additional site characteristics, see Serapiglia et al. (2014).

Table 3.2 Trait descriptions, abbreviations, and units, and time of measurement in years (yr) after coppice for field traits, and days after planting (dap) for greenhouse traits.

Trait	Abbreviation	Units	Field (yr)	Greenhouse (dap)
Biomass				
Total stem dry weight	SDW	g plant^{-1}	-	77
Total leaf dry weight	LDW	g plant^{-1}	-	77
Total root dry weight	RDW	g plant^{-1}	-	77
Stem				
Plot height	HT	m	0, 1, 2	14 − 77
Total stem length	TSL	cm	-	14 − 77
Mean stem length	MSL	cm	-	14 − 77
Stem number	STNo	#	0, 1, 2	14 − 77
Stem diameter	DIA	mm	1, 2	77
Stem area	SA	cm^2	1, 2	77
Stem volume	VOL	cm^3	1, 2	77
Stem mass	DVOL	g	1, 2	77
Axial stem number	ASN	#	-	35, 70
Foliar				
Leaf area	LFA	cm^2	1, 2	70
Leaf length	LFL	cm	1, 2	70
Leaf width	LFW	cm	1, 2	70
Leaf perimeter	LFP	cm	1, 2	70
Leaf shape factor	LFF	0−1	1, 2	70
Leaf aspect ratio	LFR		1, 2	70
Leaf dry weight	LFDW	g	1, 2	70
Specific leaf area	SLA	cm^2 g^{-1}	1, 2	70
Architecture				
Crown diameter	CDIA	cm	0, 1, 2	-
Crown form	FORM	degrees °	0, 1, 2	-
Chemical Composition				
Hemicellulose content	HCL	%	2	-
Cellulose content	CLS	%	2	-
Lignin content	LIG	%	2	-
Ash content	ASH	%	2	-
Wood density	DEN	g cm^{-3}	1, 2	-
Physiology				
SPAD	SPAD	SPAD units	0, 1, 2	14, 42, 70
Relative growth rate	RGR	cm d^{-1}	0, 1.1, 1.2, 2	14 − 77
Root electrical capacitance	REC	nF	-	77
Stem color	STC	(0, 1, 2)	0	-
Phenology				
Vegetative phenology	PHE	0−5	-	7, 9, 11, 13, 15
Pathology				
Rust severity	RUST	%	1, 2	-
Japanese beetle severity	JB	%	0, 1, 2	-
Willow leaf beetle severity	WLB	%	0, 1	-
Leaf sawfly larvae severity	SF	%	0, 1	-

3.3.4 Determination of Ploidy Level

The relative DNA content (2C-value in pg) of family parents and progeny was determined by flow cytometric analysis using young leaf material harvested from actively growing shoots in greenhouse conditions. Analysis of 50 mg of mature leaf tissue from parental genotypes and selected progeny was performed at the Flow Cytometry and Imaging Core Laboratory at Virginia Mason Research Center in Seattle, WA. A minimum of four replicates of all samples were independently assessed using either the diploid *S. purpurea* female genotype 94006 or the diploid *S. purpurea* male genotype 94001, and the tetraploid *S. miyabeana* female genotype 01-200-006 or the tetraploid *S. miyabeana* male genotype 01-200-003 as internal standards. Diploid and tetraploid parent genotypes from multiple runs were averaged and then divided by the value of the check for that run. This factor was then multiplied by each sample value within the same run as the check. When a genotype was analyzed more than once, the pg $2C^{-1}$ values were averaged.

3.3.5 Greenhouse Traits

Primary stems were defined as those emerging from dormant axillary buds and ≥ 6 cm in length. Secondary stems were defined as emerging from axillary buds on the primary stem from current-season growth (sylleptic) and were counted as the total number of secondary branches within each plot with a PHE ≥ 3. The length of each stem per plant was measured from the proximal base of the primary stem to the distal inner-whorl of the leaf primordia. The sum of stem lengths for each plot was considered to be the total stem length (TSL) and mean stem length (MSL) per plant was the mean of individual stem lengths. Starting 14 DAP to 70 DAP, all primary stems were measured within each plot once a week, totaling 9 unique time points.

The diameter of each primary stem within a plot (SDIA) was measured at the base the final week of the study using a digital caliper. Stem diameter measurements were used to calculate the area (SA) of all stems >20 cm in length. Sum of stem area per plant (SSA) was calculated by summing individual SA per plant, then modeled as a cone, it was multiplied by 1/3 plot height to obtain an estimate of total stem volume (VOL).

In order to predict root biomass (g plant^{-1}), root electrical capacitance (REC, nF) was measured according to the protocol described in (Carlson and Smart, 2016; see Chapter 2). Root biomass was harvested from a subset of 20 pots (2-3 progeny individuals per family) that were selected from the 416 pots as representing a distribution of capacitance readings, ranging from 70.5 nF to 283.8 nF. To assure the retrieval of fine root hairs, potting mix was washed from roots by first soaking root balls in water for 12 h without their pots then rinsing them by hand repeatedly and decanting into 2.4 mm and 1.0 mm aluminum test sieves. Root samples were considered appropriate for dry weight analysis when root biomass was visually free of debris. Root biomass and cuttings were separately dried in an oven and weighed.

Leaf area (LFA) was determined using a portable leaf area meter (Model No. CI-203, CID Inc., Camas, WA). A representative leaf from each plot was scanned, then excised, dried to constant weight at 65°C, and weighed to obtain leaf dry weight (LFDW). Leaf petioles were excluded from LFA and LFDW measurements. Specific leaf area (SLA) was calculated as the ratio of the total leaf area (cm^2) and dry weight (g). Leaf aspect ratio (LFR) is the ratio of the leaf length to its maximum width. Leaf shape factor (LFF, ratio of leaf area to the leaf perimeter), was corrected so that the shape factor of a circle is equal to one, i.e., $4\pi(LFA/LFP^2)$. Stem and leaf biomass was harvested separately from each plot, dried in an oven to constant weight at 65°C, then weighed to determine total stem dry weight (SDW) and total leaf dry weight (LDW).

Starting approximately 7 d after planting (DAP), the vegetative phenology stage (PHE) of each plot was scored at 7, 9, 11, 13, and 15 DAP. Vegetative phenology was scored as six stages described as: stage (0) dormant axillary buds are tightly closed and covered by bud scales; (1) axillary buds begin to swell and change color; (2) generative bud burst with visible leaves; (3) leaves emerge and begin to unfold; (4) unfolded leaves begin expanding; and (5) at least two leaves are fully expanded. To provide an indirect measurement of leaf chlorophyll and nitrogen content, leaves were measured using a Minolta SPAD 502 Chlorophyll Meter (Spectrum Technologies, Inc., Aurora, IL). Four fully-expanded leaves sampled from the upper 25% of the canopy were measured and averaged for each plant at 14, 42, and 70 DAP. Individual sex was determined visually by forcing 2-3 dormant whips to flower in greenhouse conditions.

3.3.6 Field Traits

During the dormant period after each growing season, diameters (DIA, cm) of stems ≥5 mm were measured at 30 cm from the base of the plant using Masser Racal 500 digital caliper (Masser, Rovaniemi, Finland) and stem number was counted for each plant. Total stem area (SA, cm^2) per plant was also calculated using the stem diameter values. Maximum stem height (HT, m) of every plot was recorded using a measuring rod (Crain Enterprises, Inc., Mound City, IL).

Physical and chemical wood properties were measured for four replicates. Stem segment samples were collected in the dormant period after each growing season using sampling methods previously described (Liu et al., 2015) and were stored frozen at -4°C until they were processed. The specific gravity of each sample was measured by volumetric displacement (TST om-06, 2006). In 2014, a modified method of measuring specific gravity was used where the volume of water displaced was weighed for added precision. Following specific gravity determination, stem segments were oven-dried at 65°C to a constant weight and then rough milled to a 5 mm particle

size with a Retch SM300 cutting mill (Retch, Haa, Germany) and were further comminuted to <0.5 mm particle size by fine milling with the IKA MF 10.1 knife mill (IKA, Wilmington, NC) for compositional analysis. Approximately 20 mg of each milled stem sample was analyzed with a Thermogravimetric Analyzer (TGA) Q500 instrument and Universal Analysis 2000 version 4.5A software (TA Instruments, New Castle, DE), as previously described (Serapiglia et al., 2009). Hemicellulose, cellulose, lignin, and ash content were determined as a percentage of total dry biomass for each sample, as previously described in Serapiglia et al. (2014).

At the end of the second growing season, crown diameter (CDIA, cm) was measured using modified Haglöf Mantax forestry calipers (Haglöf Sweden AB, Långsele, Sweden). Stool diameters were measured at 15.24 cm (6 in) above the soil, which is the average height of a shrub willow harvester. Crown form (FORM, degrees °) was calculated by multiplying the *arctangent2* of one-half CDIA and the fixed distance at which CDIA was measured (15.24 cm) by 180/π, to obtain the angle of the stem branching relative to the soil.

Percent rust severity (RUST, %) was visually scored by Chris Smart or Chase Crowell for each plot based on total leaf area infected. In 2017, maximum RUST scores in the association population were capped at 50% due to disease-related defoliation of heavily infected shrubs. Disease ratings were completed within a biologically relevant time period within and among field trials.

3.3.7 Statistical Analysis

All statistical analyses were performed within the statistical computing environment, R (R Core Team, 2015). For quantitative traits listed in Table 2.2, a Shapiro-Wilk test was conducted to detect a significant departure from normality. For non-normal data, the *boxcox* function was used to maximize the Shapiro-Wilk W statistic by computing log-likelihoods for

the parameter (λ) of the Box-Cox power transformation, such that either a single-parameter $(y^{\lambda} - 1)/\lambda$ or two-parameter $[(y - \lambda_2)^{\lambda_1} - 1]/\lambda_1$ power transformation was applied.

For repeated measurements of quantitative traits HT, MSL, and TSL, growth rates were determined using Gompertz 3-parameter function: $ce^{-e^{-a(t-b)}}$, whereas ordinal PHE and PSN growth rates were determined using the following 3-parameter logistic function: $c/(1 + e^{-a(t-b)})$, where, a is the growth rate, b is the inflection point, c is the asymptote, and t is time (in DAP).

Tests for association between binary and quantitative traits were done using Pearson's product moment correlation coefficient (r^2) at a confidence level of 95%. Correlations between ordinal and quantitative or binary trait pairs were tested using Spearman's rank correlation coefficient, whereas Kendall's rank correlation was used to test ordinal trait pairs. To correct for multiple comparisons, a Bonferroni correction was applied by multiplying *P*-values by the number of pairwise comparisons. Genotypes were divided based on ploidy, sex, and pedigree. To test whether two sample distributions differ, Wilcoxon Rank Sum (Mann-Whitney) test was used. When significant differences were observed, treatment comparisons were performed using Tukey's Range test.

Variance components for the greenhouse trial were estimated with *lmer* in the package lme4 (Bates et al., 2015) using the Restricted Maximum Likelihood (REML) method, for the following model:

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$$

where y_{ijk} is the observed value, μ is the overall mean, α_i is the effect of genotype i, β_j is the effect of block j, and ε_{ijk} is the random error, which is assumed independent and identically distributed.

Field trial dimensions were 388.6 m × 36.6 m (Figure 3.1), which introduced spatial variation not easily accountable by block alone. Thus, to account for spatial variation in the field trial, following the approach outlined in Velazco et al. (2017), spatial trends (row and column) in the field trial were modeled as two-dimensional Penalized (P)-splines, using *SpATS* and *SAP* functions (n.seg = (16, 64), tolerance = 1×10^{-6}) in the SpATS package (Rodríguez-Álvarez et al., 2015, Rodríguez-Álvarez et al., 2017).

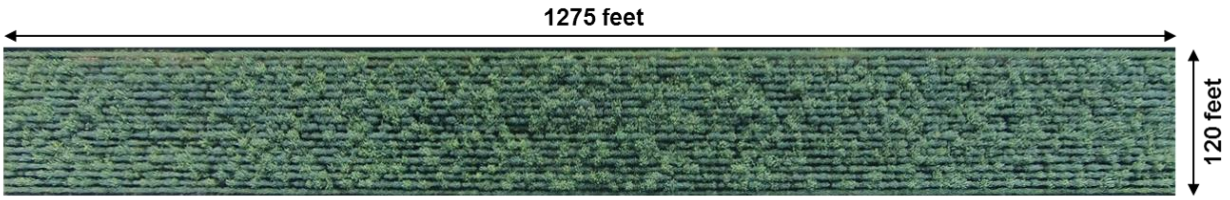

Figure 3.1 Aerial image of the CN009 field trial located at Cornell AgriTech (Geneva, NY).

The conditional means of random effects were extracted from the above linear mixed model in order to provide the relative phenotypic contributions of the male and female parents. The percent deviation of the F_1 progeny trait mean relative to the midparent value (MPH) was calculated as: $[(F_1 - MP)/MP] \times 100$, where F_1 is the progeny mean and MP is the geometric mean of the female parent and the male parent. In this case, genotype (clone) was fixed in the analysis, so MPH represents deviations from these estimates. Fisher's Exact test was used to classify phenotypic expression into modes of inheritance ($P < 0.05$) by comparing the deviation of the F_1 to both the female and male parents. For the reason that the male parent of the intraspecific tetraploid F_1 *S. miyabeana* family 425 ('SX64') was not present, only the deviation of the F_1 from the female parent 01-200-006 was reported.

3.4 Results

3.4.1 Harvestable Biomass in the Greenhouse Trial

Of the 104 genotypes harvested in this trial, total aboveground biomass (AGB, g) ranged from 119.6 g to 51.1 g (Figure 3.2). The greatest yielding genotype was a triploid hybrid, *S. viminalis* × *S. miyabeana* 12X-423-043, and the least yielding genotype was a diploid *S. purpurea*, 10X-082-078. The greatest mean AGB was from the *S. viminalis* × *S. miyabeana* triploid family 423 (100.3 ± 3.8) and the *S. miyabeana* × *S. viminalis* triploid family 430 (98.7 ± 3.4), followed by the *S. viminalis* × *S. purpurea* diploid family 421 (92.3 ± 3.8). All other families were not significantly different from one another, with a family mean AGB ranging from 78.9 to 82.5 g.

The mean AGB of triploids (93.8 ±1.6) was significantly greater (Wilcoxon $P < 0.001$) than the mean AGB of diploids (82.6 ±1.3) and tetraploids (79.8 ±2.6). While the triploid families 423 and 430 did show higher AGB than both diploid and tetraploid parents, AGB did not exceed the midparent for family 415 (Figure 3.3). Similarly, belowground biomass (RDW, g), which was estimated using the REC method described in Carlson and Smart (2016), ranged from 11.1 g to 42.5 g, and showed that the triploid family 430 had accumulated the greatest RDW, followed by families 423 and 421, with the lowest mean family REC values from the *S. purpurea* × *S. miyabeana* family 415. The mean RDW of triploids (26.6 ±0.8) was significantly greater (Wilcoxon $P < 0.001$) than that of diploids (23.1 ±0.6) and tetraploids (22.2 ±0.9), there was no significant difference between diploids and tetraploids (Wilcoxon, $P = 0.57$).

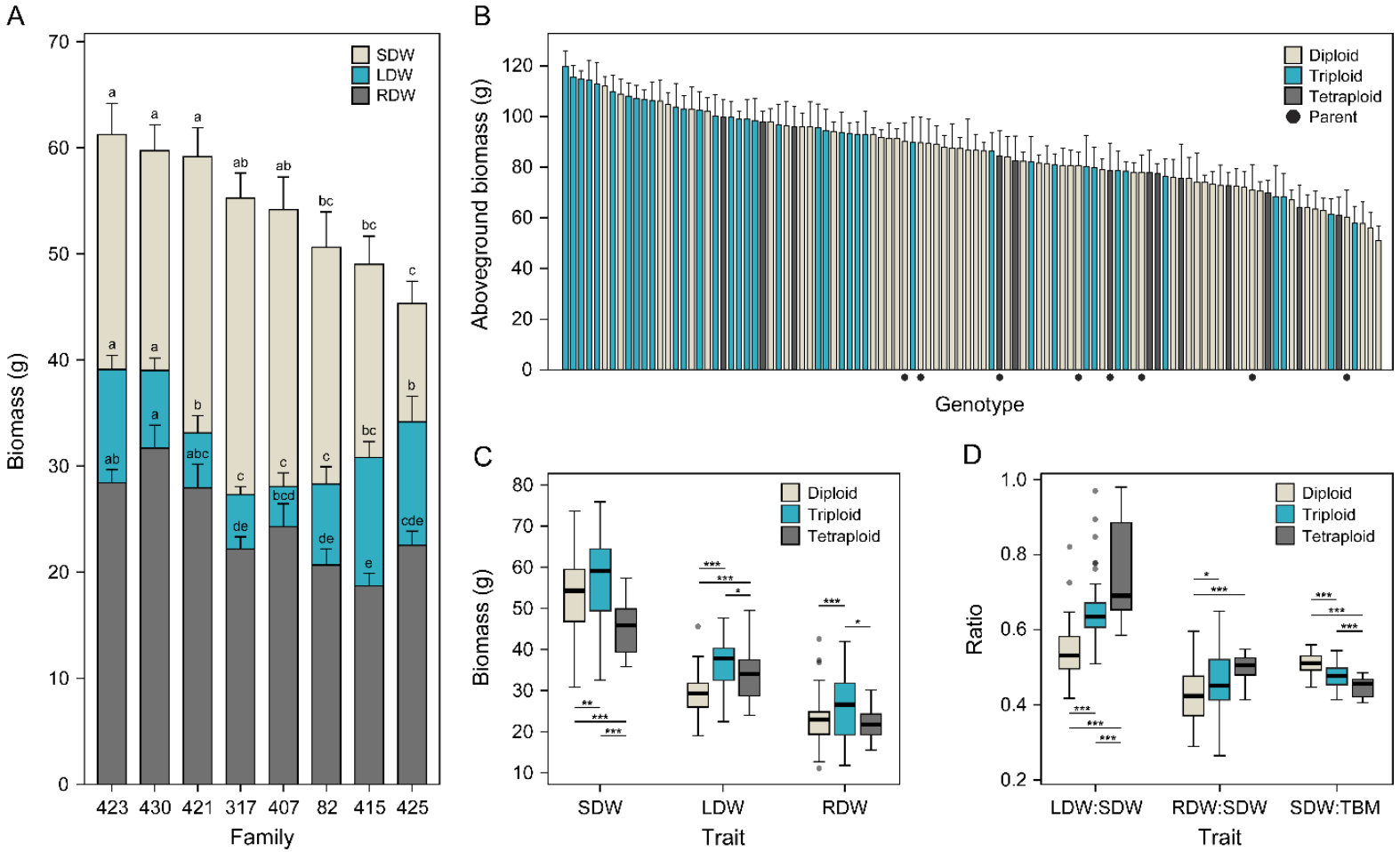

Figure 3.2 Harvested biomass by family, genotype, and ploidy. Overlain **(A)** barplots of mean (±SE) SDW, LDW, and RDW biomass by family. For each trait, letters above bars represent significant differences by family according to Tukey's HSD groupings (α = 0.05). The **(B)** aboveground biomass means (±SE) all parent and progeny genotypes (n = 104) are shown in descending order. Bars are filled according to the ploidy of each genotype and filled circles below bars specify the location of the parents. Boxplot distributions depict the median and interquartile range (IQR ±1.5) of **(C)** biomass dry wt. **(D)** and biomass dry wt. ratios by ploidy, where asterisks *,**,*** denote significant differences at a Wilcoxon $P < 0.05$, < 0.01, and < 0.001, respectively.

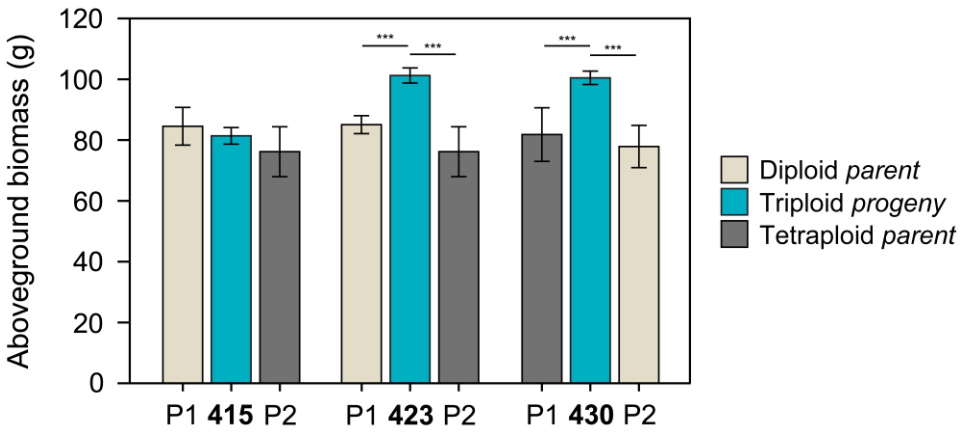

Figure 3.3 Total aboveground biomass averages of triploid families and respective female (P1) and male (P2) parents. Bars are colored according to ploidy-level (see legend). Asterisks *** above bars denote significant differences ($P < 0.001$).

Biomass ratios LDW SDW^{-1} and RDW SDW^{-1} for tetraploid genotypes were significantly greater than those of triploids and diploids (Wilcoxon $P < 0.001$), whereas diploids had significantly lower LDW SDW^{-1} and RDW SDW^{-1} ratios compared to both triploids and tetraploids (Wilcoxon $P < 0.001$) (Figure 3.2D). The tetraploid *S. miyabeana* family 425 had the greatest LDW SDW^{-1} ratio (0.76 ±0.02) and the triploid *S. miyabeana* × *S. viminalis* triploid family 430 had the greatest RDW SDW^{-1} ratio (0.53 ±0.02). Further, the diploid F_2 *S. purpurea* family 317 had the lowest LDW SDW^{-1} ratio (0.50 ±0.01) and the triploid *S. purpurea* × *S. miyabeana* family 415 had the lowest RDW SDW^{-1} ratio (0.38 ±0.02), yet was not significantly different from both intraspecific F_1 and F_2 *S. purpurea* families. Subsequently, the ratio of SDW to the total biomass (TBM) showed the converse, which suggests that diploids placed considerably less energy into leaf biomass production (28.9 ±0.4) compared to triploids (36.6 ±0.6, Wilcoxon $P < 0.001$) and tetraploids (34.4 ±1.4, Wilcoxon $P < 0.001$).

3.4.2 Biomass-Related Stem Growth in the Greenhouse

Growth measurements taken at the end of the study included the length and diameter of every stem, as well as primary and axial stem number. The greatest plot HT after 84 DAP in the

greenhouse was observed for the triploid hybrid, *S. viminalis* × *S. miyabeana* 12X-423-034 (2.1 ±0.05), whereas the genotype with the lowest HT was the intraspecific tetraploid, *S. miyabeana* 12X-425-106 (1.2 ±0.09). Total stem length (TSL) was calculated as the sum of the length of each shoot >20 cm per plot, and ranged from 6.03 m to 1.95 m. While TSL of diploids (3.89 ±0.08) and triploids (3.88 ±0.10) was not significantly different (Wilcoxon $P = 0.85$) at 84 DAP, tetraploids showed significantly less TSL (3.47 ±0.13) than diploids (Wilcoxon $P = 0.02$) and triploids (Wilcoxon $P = 0.01$). The lower TSL values for tetraploids may be explained by lower PSN and ASN after 84 DAP, thus leading to fewer stems and lower TSL than diploids and triploids. Although triploids had a greater MSL (1.62 ±0.02) at 84 DAP (Wilcoxon $P < 0.001$), diploids (1.44 ±0.02) and tetraploids (1.49 ±0.04) were not significantly different (Wilcoxon $P = 0.11$). The sum of DIA sampled were used to calculate MDIA and sum stem area (SSA) for each plot. The MSA ranged from 0.37 to 1.41 cm^2 and SSA ranged from 1.29 to 2.95 cm^3.

3.4.3 Physiological and Foliar Traits

For each plot, three SPAD readings were taken at 14, 42, and 70 DAP to assess the chlorophyll content or nitrogen status of fully-expanded leaves from the upper 30 cm of the canopy. A significant interaction was identified for SPAD by time and ploidy as well as time and family. While the initial reading at 14 DAP showed marginally greater SPAD values for triploids and tetraploids, the opposite trend was found for later readings (Figure 3.4). Diploid genotypes showed significantly greater SPAD readings at 42 DAP than triploids and tetraploids, which were not significantly different. By 70 DAP, differences in SPAD readings by ploidy were not as great as those taken at 42 DAP. This was also observed for SPAD readings taken in the field trial, where increased ploidy was inversely correlated with greater SPAD readings.

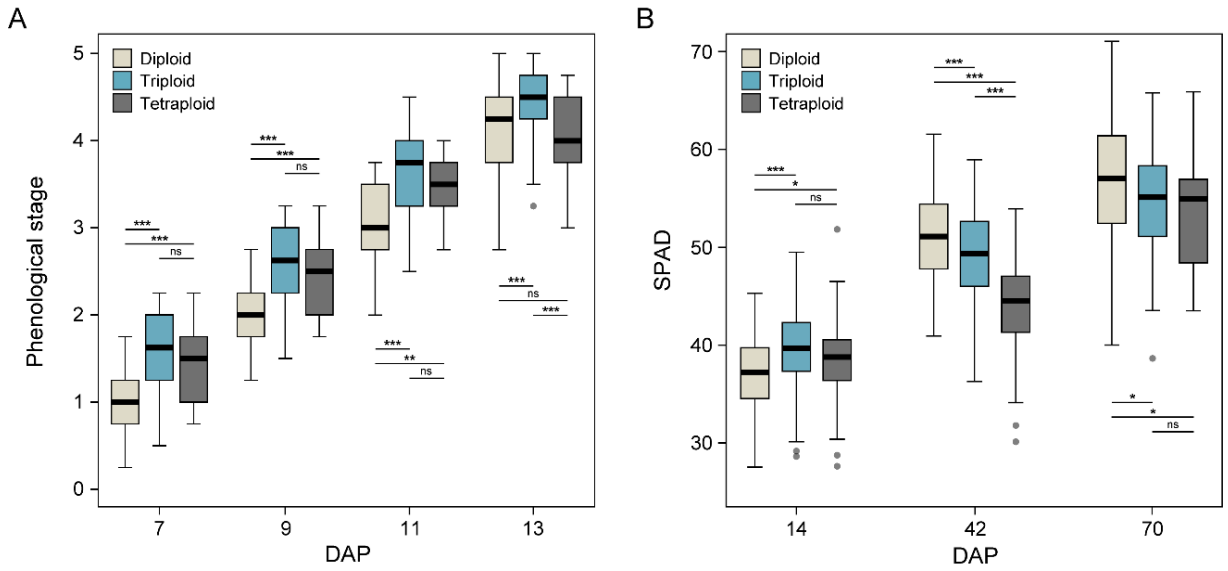

Figure 3.4 Repeated physiological measurements by ploidy. Boxplot distributions depict the median and interquartile range (IQR ±1.5) of **(A)** phenological stages (PHE) at 7, 9, 11, and 13 days after planting (DAP) as well as **(B)** SPAD values by ploidy. Asterisks *,**,*** above or below boxplots of diploids (beige), triploids (cyan), and tetraploids (dark grey) denote significant differences at a Wilcoxon $P < 0.05$, < 0.01, and < 0.001, respectively. The 15 DAP stage is not shown because there were no significant differences by ploidy.

Leaf measurements obtained at 11 weeks included leaf area (LFA), length (LFL), width (LFW), perimeter (LFP), ratio (LFR), and dry weight (LFDW). Of all the genotypes in the trial, the intraspecific tetraploid, *S. miyabeana* 13X-425-110, had the greatest LFA (40.0 ±7.5), whereas the diploid hybrid, *S. purpurea* × *S. viminalis* 11X-407-087, had the lowest LFA (8.9 ±1.6). On a family-level, the intraspecific *S. miyabeana* family 425 had the greatest LFA (22.0 ±1.2) and the intraspecific F_1 *S. purpurea* family 82 had the lowest LFA (13.9 ±0.5). Average LFL ranged from 17.5 cm to 6.6 cm. Among the top 50th percentile for the leaf dimensions LFA, LFL, and LFW, nearly all genotypes had either the *S. viminalis* female 07-MBG-5027 or the *S. viminalis* male 'Jorr' as a parent.

3.4.4 Midparent Heterosis and Inheritance in the Greenhouse

Relative to diploid and tetraploid genotypes, these data show that triploids exhibit greater levels of heterosis for biomass yield and correlated growth traits, especially for crosses made

between *Salix* Sections (Figure 3.5, Figure 3.6). The interspecific diploid family 421 and interspecific triploid families 423 and 430 have *S. viminalis* as one of the parents, and all showed high-levels of MPH (%) for total SDW and RDW, as well as RGR and early TSL and PHE measurements. However, MPH for total biomass or stem growth measurements were not observed in interspecific *S. purpurea* × *S. viminalis* family 407 or the intraspecific *S. purpurea* or *S. miyabeana* F_1 families.

The $\log_2$ difference of the respective female and male parent from the family progeny was determined in order to assess global inheritance patterns for all biomass-related traits in each family (Figure 3.7). The diploid F_2 *S. purpurea* family 317 showed the most conserved inheritance for all traits, whereas the diploid F_1 *S. purpurea* family 82 showed greater levels of both P1- and P2-dominant as well as underdominant inheritance. The reciprocal interspecific diploid families 407 and 421 showed strong patterns of dominance, almost exclusively in the direction of the *S. viminalis* parent species. While stem traits for the triploid *S. purpurea* × *S. miyabeana* family 415 primarily displayed P2-dominant inheritance, foliar traits within the family reflected more conserved or additive inheritance. The triploid *S. viminalis* × *S. miyabeana* family 423 showed strong P1-dominant inheritance for all traits, comparable to the diploid *S. viminalis* × *S. purpurea* family 421. Although many traits in family 423 showed P1-dominant inheritance, conserved or additive inheritance for foliar traits were predominant, compared to that of stem traits in the same family. Unlike triploid family 423, the triploid *S. miyabeana* × *S. viminalis* family 430 had the greatest number of transgressive traits compared to all other families, and showed nearly equal P1- and P2-dominant inheritance patterns.

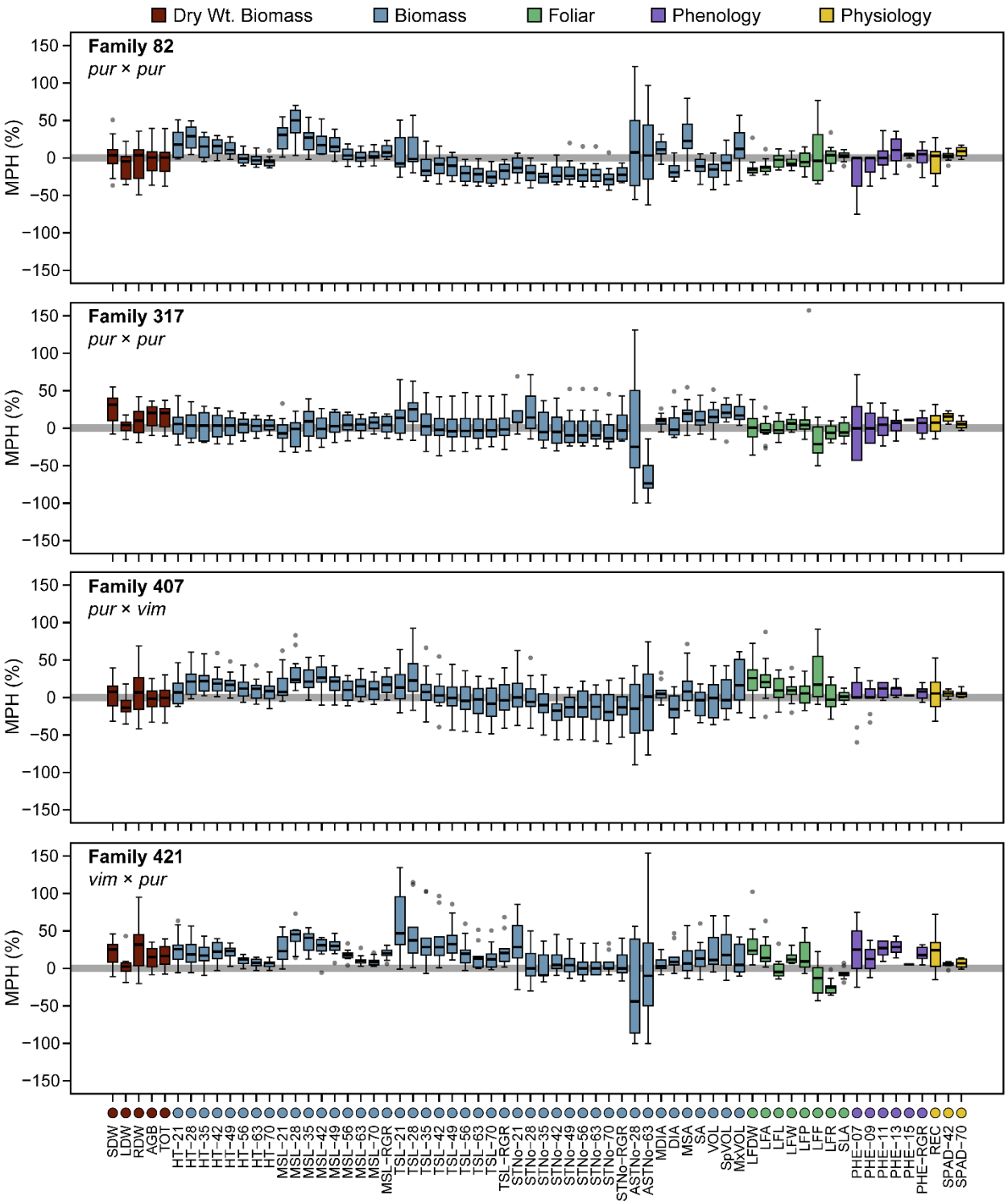

Figure 3.5 Midparent heterosis (MPH %) for greenhouse collected traits in diploids. Boxplot distributions are shown as the percent deviation of the hybrid from the midparent, and depict the median and interquartile range (IQR ±1.5) of MPH for each trait by family, which are filled according to the legend above panels.

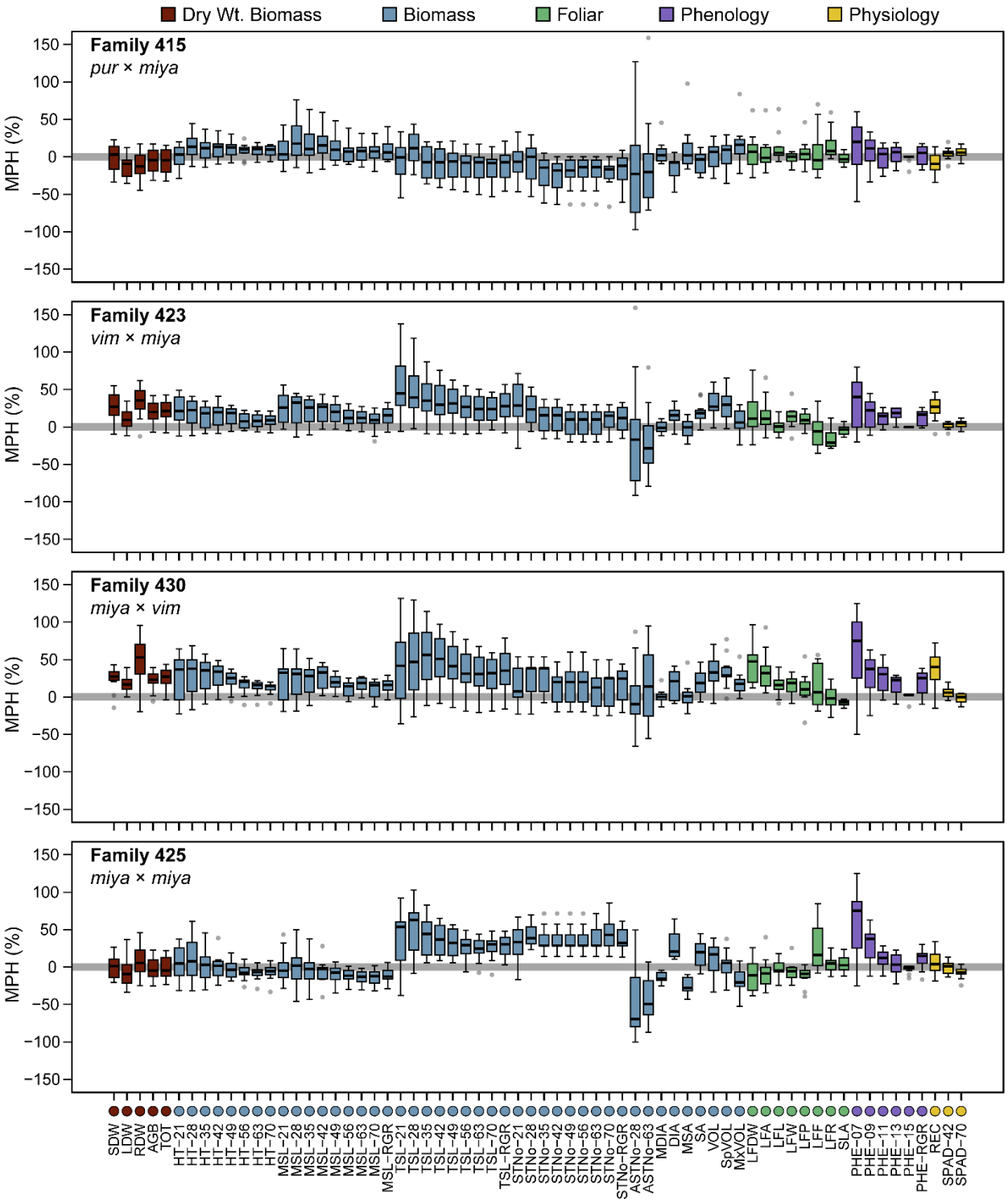

Figure 3.6 Midparent heterosis (MPH %) for greenhouse collected traits in triploids and tetraploids. Boxplot distributions are shown as the percent deviation of the hybrid from the midparent, and depict the median and interquartile range (IQR ±1.5) of MPH for each trait by family, which are filled according to the legend above panels.

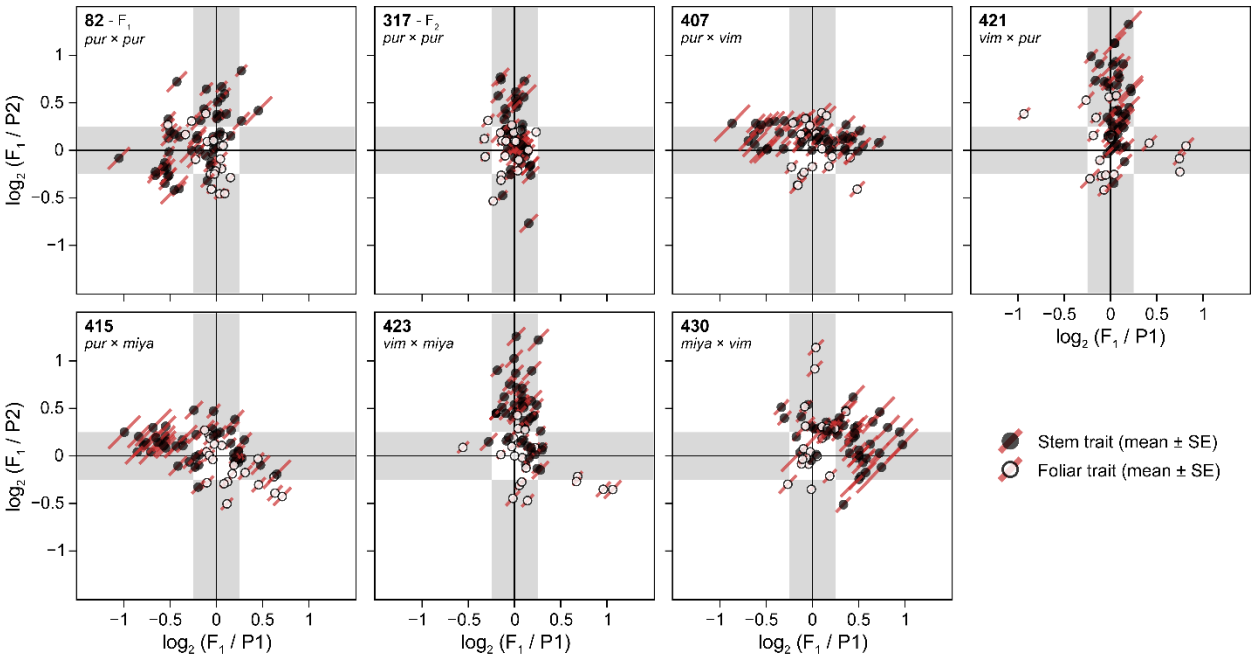

Figure 3.7 Inheritance patterns growth traits among hybrid shrub willow families. Each point represents a single trait plotted as the $\log_2$ difference of the progeny from the female ($\log_2$ (F_1/P1), horizontal axis) and male ($\log_2$ (F_1/P2), vertical axis) parents. Stem traits are characterized by filled black points and foliar traits, by filled white points. Red lines passing through each point represents the corresponding ± standard error of the family mean.

3.4.5 Multivariate Analysis

Traits that were non-informative in biomass yield predictions, but were correlated with informative predictors may still prove to be of relative importance, particularly when assayed in additional environments or pedigrees. While many of the traits listed in Table 3.2 were strongly correlated to those important for biomass yield, some pairs tended to be more autocorrelated (Figure 3.8), as they were repeated measurements or components of the same trait. The greenhouse-collected traits showing the highest correlations with SDW were VOL (r^2 = 0.73), REC (r^2 = 0.78), and LDW (r = 0.62); all grouping very closely in Figure 3.8. Following these traits, later stem length measurements were most correlated to SDW, whereas, PHE at 11 and 13 DAP and ASN at 56 DAP were most correlated with LDW. Although weakly, both SDW and LDW had inverse relationships with the inflection point of HT.

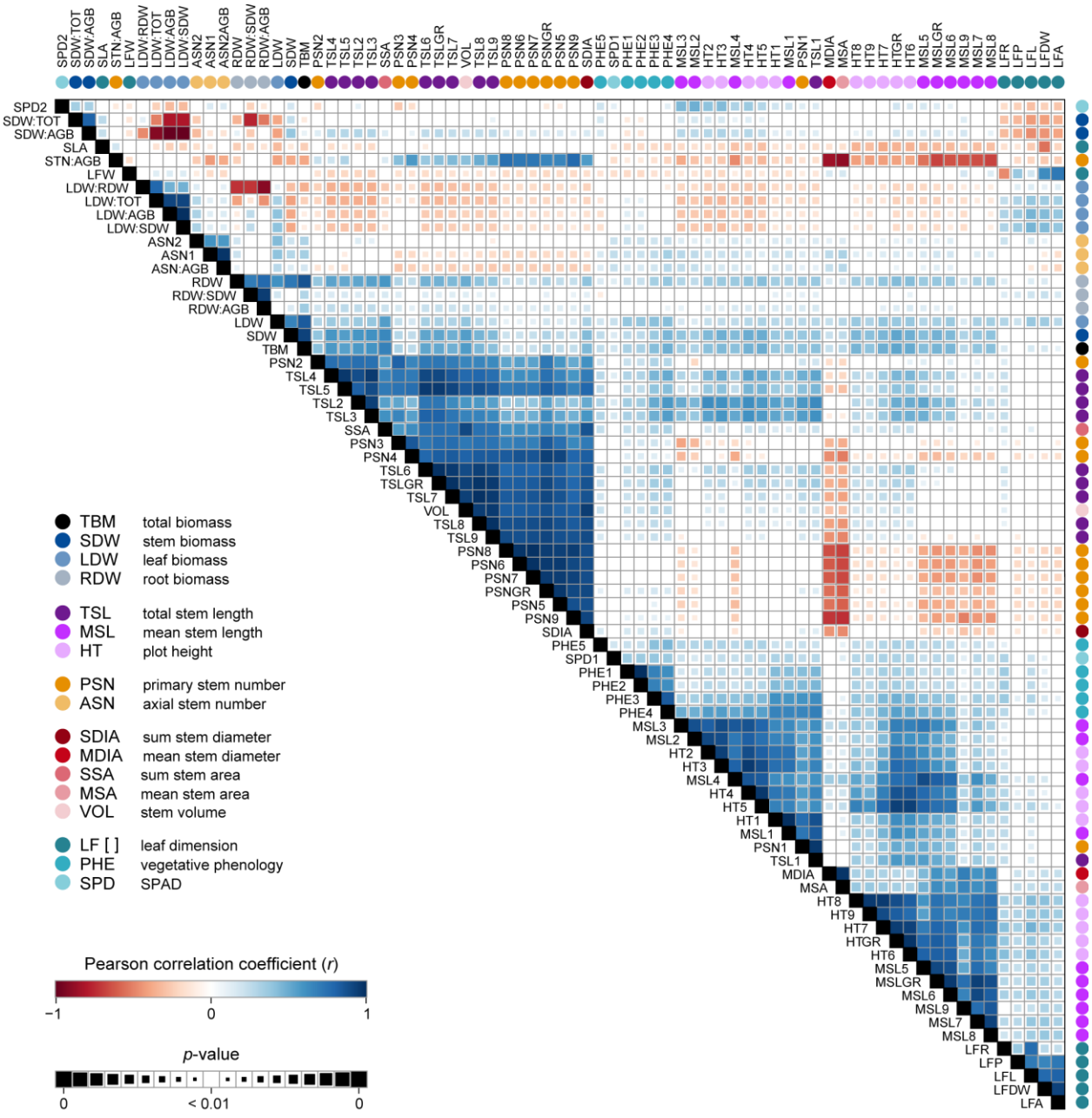

Figure 3.8 Pairwise correlations of biomass traits collected in the greenhouse. The correlation matrix was re-ordered by hierarchical clustering using the *average* distance method. Beneath trait abbreviations in the first row and in the final column, filled-circles highlight major trait classes detailed in the legend. Increasing in intensity with higher absolute Pearson correlation coefficients (*r*), positive correlations are illustrated by filled blue squares and negative correlations by filled-red squares. Non-significant correlations ($p > 0.01$) were left blank. Significance levels (*p*-values) were used to scale the area of each square, such that smaller squares represent correlations with lower significance and larger squares represent those showing more significance.

Phenological stages (PHE) recorded at 11 and 13 DAP in the greenhouse were strongly correlated with SPAD at 14 DAP, LFR, LFL, TSL at 14 and 21 DAP, and the growth rates of HT, MSL, TSL, but not STN. Early PHE measurements at 7 and 9 DAP were positively correlated with CVAL and LFR. At 13 and 15 DAP, PHE was only weakly correlated with SPAD at 42 DAP. All PHE measurements were inversely correlated with LFW, SLA, and the ratio STN AGB^{-1}. Leaf dimensions LFL, LFA, and LFP, as well as the biomass ratio LDW SDW^{-1} were strongly and positively correlated ($r^2 > 0.5$) with CVAL, whereas both SPAD at 42 DAP and SLA were inversely correlated ($r^2 < -0.5$). Besides showing positive correlations to the ratios SDW TOT^{-1} and STN AGB^{-1}, SLA was the solitary trait to be inversely correlated with nearly all growth traits. Along with SLA, SDW TOT^{-1} and STN AGB^{-1} were most negatively correlated with leaf dimensions, as well as SPAD at 42 DAP and STN measurements. Root electrical capacitance (REC), SSA, TSL at 28 DAP, and PHE at 11 and 13 DAP were highly correlated with SDW, and have been shown to account for a large proportion of the variance ($R^2 = 0.69$) in multiple linear regression (Carlson and Smart, 2016; see Chapter 2).

For field-collected biomass traits, all were highly correlated for both years (Figure 3.9). However, foliar traits were positively, but more weakly correlated between years, besides leaf ratio (LFR) ($r^2 = 0.65$, $p < 0.001$) and leaf shape factor (LFF) ($r^2 = 0.35$, $p < 0.001$). Crown form (FORM) measurements for all three years measured were most inversely correlated with biomass stem growth traits (e.g., SA, HT, VOL, and STNo) as well as wood density (DEN). Wood chemical composition traits were also highly correlated, whereby LIG and ASH were inversely correlated with CLS and HCL. Wood density (DEN) was only positively correlated with CLS, but inversely correlated with HCL and ASH. Overall, individuals with higher ploidy-levels tended to have higher ASH and lower HCL, compared with diploids (Figure 3.10).

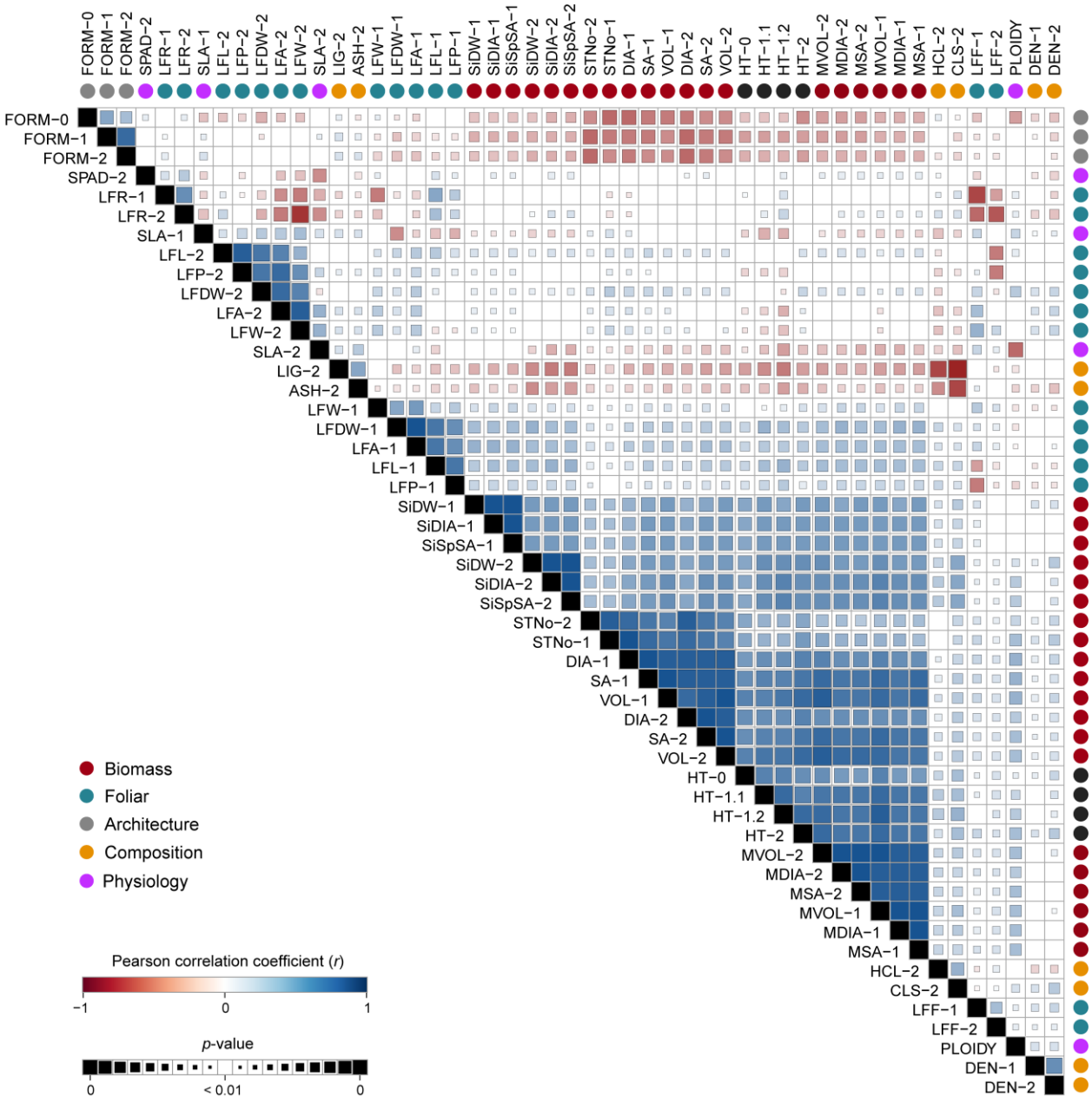

Figure 3.9 Pairwise correlations of biomass traits collected in the field. The correlation matrix was re-ordered by hierarchical clustering using the *average* distance method. Beneath trait abbreviations in the first row and in the final column, filled-circles highlight major trait classes detailed in the legend. Increasing in intensity with higher absolute Pearson correlation coefficients (*r*), positive correlations are illustrated by filled blue squares and negative correlations by filled-red squares. Non-significant correlations ($p > 0.01$) were left blank. Significance levels (*p*-values) were used to scale the area of each square, such that smaller squares represent correlations with lower significance and larger squares represent those showing more significance.

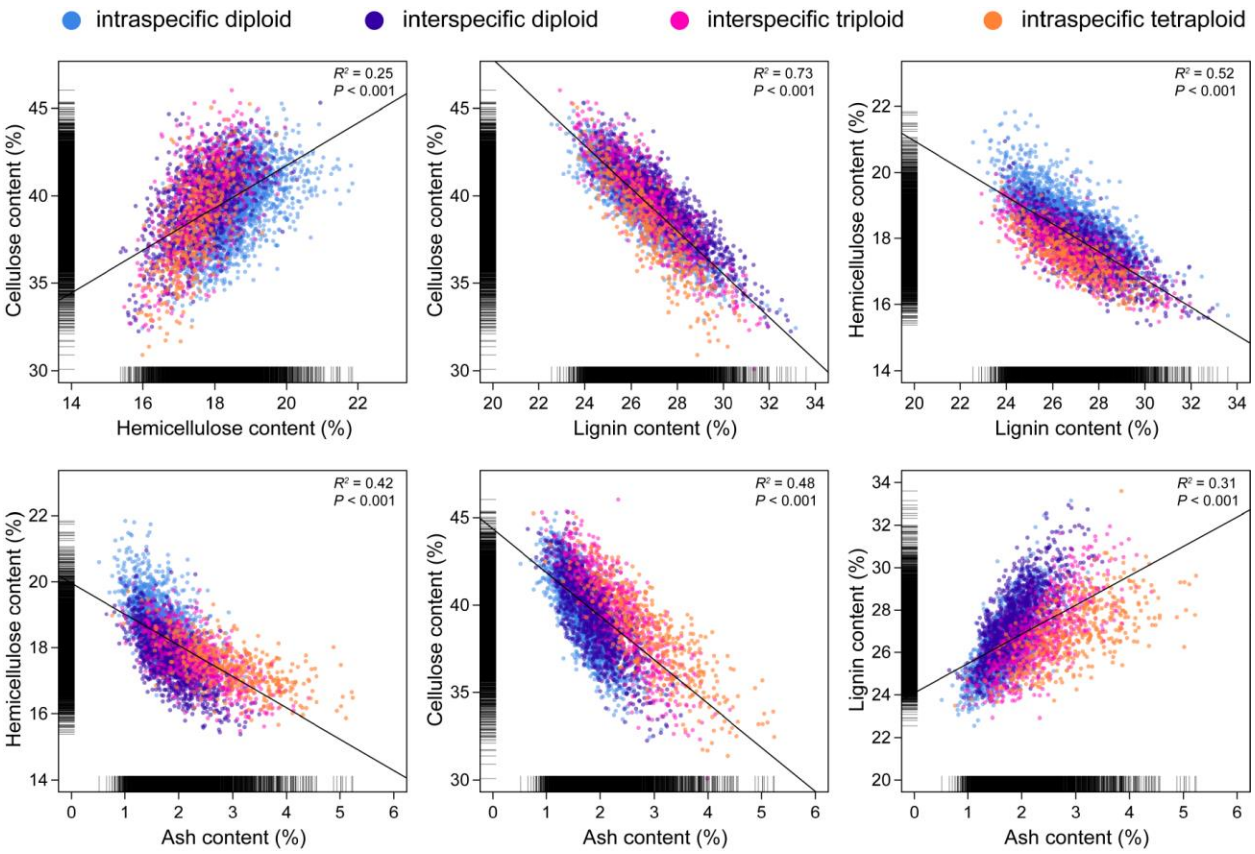

Figure 3.10 Wood chemical composition associations from second year post-coppice measurements. Points within each scatterplot are colored according to individual ploidy-level, according to the legend above figure panels.

3.4.6 Concordance of Heterosis for Common Greenhouse and Field Traits

Field-collected traits for the same families resulted in similar levels of heterosis for common traits collected in the greenhouse trial (Figures 3.11, 3.12). For instance, families 423 and 430 showed the greatest MPH for HT, SA, and VOL for both years in the field trial as well as in the greenhouse trial. While the interspecific diploid families 407 and 421 showed marginal levels of MPH for the same traits, it was not realized in the field trial as a result of a high-incidence of potato leafhopper and Japanese beetle feeding on family individuals with *S. viminalis* in their background. Yet this was not observed for triploid families with a diploid *S. viminalis* parent, which suggests a resistance effect at higher ploidies.

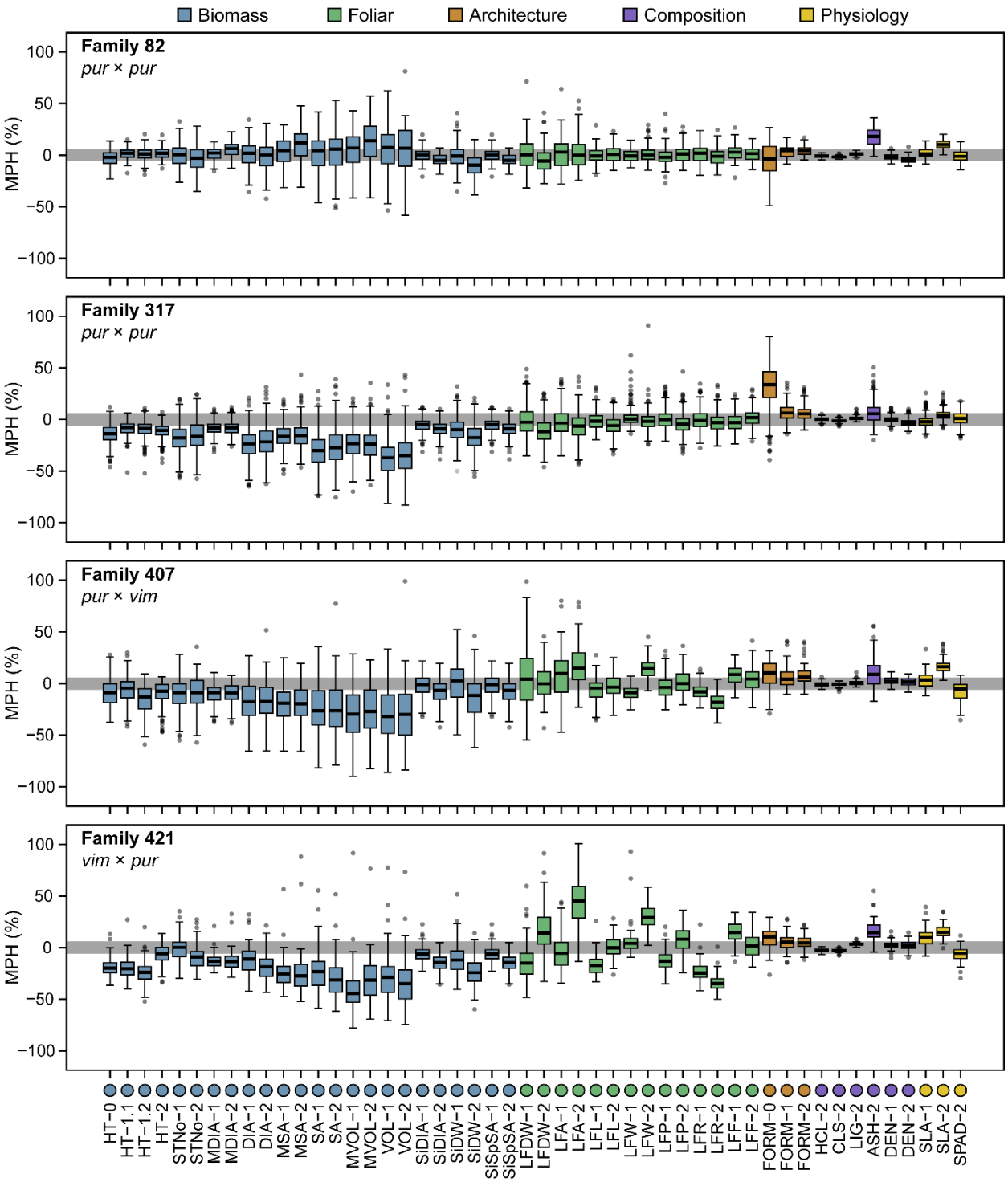

Figure 3.11 Midparent heterosis (MPH %) for field collected traits in diploids. Boxplot distributions are shown as the percent deviation of the hybrid from the midparent, and depict the median and interquartile range (IQR ±1.5) of MPH for each trait by family, which are filled according to the legend above panels.

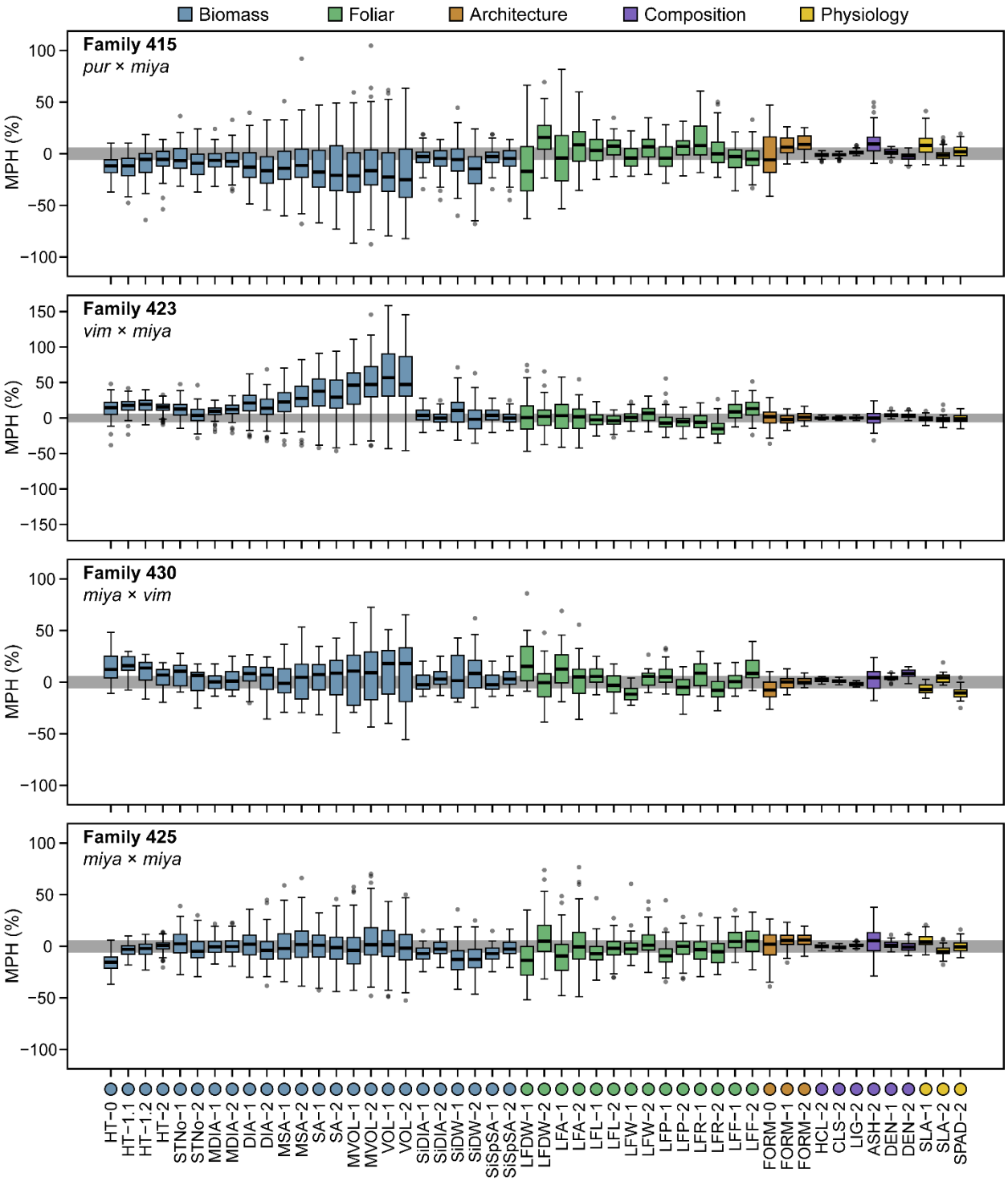

Figure 3.12 Midparent heterosis (MPH %) for field collected traits in triploids and tetraploids. Boxplot distributions are shown as the percent deviation of the hybrid from the midparent, and depict the median and interquartile range (IQR ±1.5) of MPH for each trait by family, which are filled according to the legend above panels.

Of the two years in the field trial, foliar traits were shown to be the most variable between years. This was due to differences in precipitation of the 2015 and 2016 growing seasons. In 2015 growing conditions were nearly optimal throughout the season, but 2016 suffered a long stretch of drought mid-summer. Although individuals did not dramatically differ in MPH for most trait rankings between growing seasons, foliar trait MPH variation can be assessed by SLA. However, families 421, 423, and 430, which all had an *S. viminalis* parent, did show greater MPH for foliar traits in both the greenhouse and field trials. Surprisingly, individuals, 12X-423-060 and 12X-423-110, had average LFL > 40 cm, which was nearly two-fold greater than either better parent.

3.5 Discussion

3.5.1 Leaf to Shoot Biomass Ratio

Given that all plots received the same fertilizer and application rate, it may be that the nitrogen status of leaves is concentrated to less leaf area in diploids, given the significantly greater SPAD values in diploids, compared to those of triploids and tetraploids. Further, LFA and LDW of triploid F_1 individuals showed primarily additive inheritance, and on the basis of ploidy, triploids were intermediate to diploids and tetraploids for the same traits. Yet, under controlled environmental conditions, higher ploidy tended to result in greater leaf area and biomass, but a lower leaf nitrogen status. One possibility to explain this is that polyploid willows are more efficient in the production of *cheap* leaves (Fabio et al., 2018); perhaps by focusing available nutrient resources to a rapidly emerging canopy, rather than uniformly along the stem, as is likely the case in diploids. Triploid genotypes with an intermediate LDW SDW^{-1} ratio could indicate more efficient partitioning of photoassimilate from leaves to sink organs. While this attribute would surely stimulate the rapid accumulation of biomass, diploids could benefit in

nutrient-scarce environments by sustaining growth rates, whereas higher ploidy levels would likely show an overall reduction in their growth rate.

3.5.2 Ploidy Differences in Growth Rates

Repeated measures have commonly been used to identify growth patterns in response to environmental factors or treatments. Here, the treatments under consideration were pedigree and ploidy level. While the inherent differences between factors tend to inflate the actual differences (e.g., diploids versus tetraploids), relative growth rates (RGR) act as a standardized measure of growth, and offer more impartial comparisons. Early vegetative PHE measurements showed that triploids are faster to break bud and grow at faster rates compared with diploids and tetraploids. However, over time, diploids maintained more linear growth rates compared to that of triploids, which leveled-off approximately 8 weeks after planting. This could be due to increasingly limited space in the pots of triploids, as they exhibited both greater above and below-ground biomass at the termination of the study, especially for those with *S. viminalis* as one of the parents, as described in Carlson and Smart (2016) (see Chapter 2). It may be that *S. viminalis* crosses have a higher propensity for accumulating root mass than intraspecific or interspecific crosses of *S. purpurea* and *S. miyabeana*. Using the REC phenotyping method, genetic mapping for this trait could improve our understanding of root development and response to drought among willow crosses. Marker-assisted selection (MAS) for REC could subsequently be used to improve biomass yield, in the case that field evaluations correlate well with those in the greenhouse.

2.5.3 Heterosis for Biomass-Related Growth Traits

The genus *Salix* is estimated to comprise 350 dioecious perennial species (Kuzovkina et al., 2008). As obligate outcrossers, *Salix* spp. are primarily insect pollinated (entomophilous) and

may hybridize in natural settings. The heterozygous nature of *Salix* prevents a genetically uniform F_1 and both hybrid vigor and hybrid necrosis can be represented by siblings of the same cross. Plants expressing heterosis are thought to have experienced one or more duplication events in their past (Soltis et al., 2009) and many crop plants are polyploid. With respect to modeling heterosis in interspecific hybrid plants, the body of literature has been centered on the comparison of hybrids generated from inbred or synthetic allopolyploids. Polyploidy via chromosome doubling or wide-hybridization has demonstrated the positive impact on the accumulation of biomass in hybrids, compared to progenitors (Chen 2010). Yet, in most plant crops, aneuploidy generally corresponds with hybrid necrosis (Birchler 2007), whereby sub-optimal gene dosages negatively impact protein function and metabolic homeostasis. Here, we demonstrate that triploid shrub willow hybrids, derived from diploid and tetraploid parents of different species, exhibit high levels of both dominant and transgressive phenotypic expression. Most notably, this was observed for biomass growth traits for crosses between *Salix* Sections *Helix* and *Vetrix*.

3.6 Conclusion

The results outlined here corroborate consistent findings that triploids produce greater biomass yields than their diploid or tetraploid parents. Heterosis for many of the extensive traits collected in the greenhouse also showed heterosis in the field, with consistently greater total stem volume among triploid individuals. The genetic basis of heterosis in willows is not well-understood, and further work on characterizing this phenomena will support community efforts in build a toolkit for improving this sustainable, fast-growing bioenergy crop.

3.7 REFERENCES

Argus GW. 1974. An experimental study of hybridization and pollination in Salix (willow). *Canadian Journal of Botany,* 52: 1613–619.

Argus GW. 1997. Infrageneric classification of *Salix* (*Salicaceae*) in the New World. *Systematic Botany Monographs,* 52: 1–121.

Bates D, Mächler M, Bolker B, Walker S. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software,* 67: 48.

Cameron KD, Phillips IS, Kopp RF, Volk TA, Maynard CA, Abrahamson LP, Smart LB. 2008. Quantitative genetics of traits indicative of biomass production and heterosis in 34 full-sib F_1 *Salix eriocephala* families. *BioEnergy Research,* 1: 80–90.

Carlson CH, Smart LB. 2016. Electrical capacitance as a predictor of root dry weight in shrub willow (*Salix*; Salicaceae) parents and progeny. *Applications in Plant Science,* 4: 1600031.

Fabio ES, Kemanian AR, Montes F, Miller RO, Smart LB. 2017a. A mixed model approach for evaluating yield improvements in interspecific hybrids of shrub willow, a dedicated bioenergy crop. *Industrial Crops and Products,* 96: 57–70.

Fabio ES, Volk TA, Miller RO, Serapiglia MJ, Gauch HG, Van Rees KC, Hangs RD, Amichev BY, Kuzovkina YA, Labrecque M. 2017. Genotype by environment interactions analysis of North American shrub willow yield trials confirms superior performance of triploid hybrids. *GCB Bioenergy,* 9: 445–459.

Fabio ES, Volk TA, Miller RO, Serapiglia MJ, Kemanian AR, Montes F, Kuzovkina YA, Kling GJ, Smart LB. 2017b. Contributions of environment and genotype to variation in shrub willow biomass composition. *Industrial Crops and Products,* 108: 149–161.

Grömping U. 2006. Relative Importance for Linear Regression in R: The Package relaimpo. *Journal of Statistical Software,* 17: 1-27.

Hardig TM, Brunsfeld SJ, Fritz RS, Morgan M, Orians CM. 2000. Morphological and molecular evidence for hybridization and introgression in a willow (Salix) hybrid zone. *Molecular Ecology,* 9: 9–24.

Heribert-Nilsson N. 1918. *Experimentelle studien uber variabiliat, Spaltung, Artibildung, und evoluation in der gattung Salix*, Lunds Universitats Arsskrift N. F. Avd 2.

Karp A, Hanley S, Trybush S, Macalpine W, Pei MH, Shield I. 2011. Genetic improvement of willow for bioenergy and biofuels. *Journal of Integrative Plant Biology,* 53.

Kuzovkina YA, Weih M, Romero MA, Charles J, Hust S, McIvor I, Karp A, Trybush S, Labrecque M, Teodorescu TI, Singh NB, Smart LB, Volk TA. 2008. *Salix*: Botany and Global Horticulture. *Horticultural Reviews*: John Wiley & Sons, Inc.

Lauron-Moreau A, Pitre FE, Argus GW, Labrecque M, Brouillet L. 2015. Phylogenetic relationships of American willows (*Salix* L., Salicaceae). *PLoS ONE,* 10: e0121965.

Lin J, Gibbs JP, Smart LB. 2009. Population genetic structure of native versus naturalized sympatric shrub willows (*Salix*: Salicaceae). *American Journal of Botany,* 96: 771–785.

Lindeman RH, Merenda PF, Gold RZ. 1980. Introduction to bivariate and multivariate analysis. Glenview, IL: Longman Higher Education, 1–444.

Percy DM, Argus GW, Cronk QC, Fazekas AJ, Kesanakurti PR, Burgess KS, Husband BC, Newmaster SG, Barrett SCH, Graham SW. 2014. Understanding the spectacular failure

of DNA barcoding in willows (*Salix*): Does this result from a trans-specific selective sweep? *Molecular Ecology,* 23: 4737–4756.

R Core Team. 2015. *R: A language and environment for statistical computing.* Vienna, Australia: R Foundation for Statistical Computing.

Serapiglia MJ, Gouker FE, Smart LB. 2014. Early selection of novel triploid hybrids of shrub willow with improved biomass yield relative to diploids. *BMC Plant Biology,* 14: 74.

Smart LB, Cameron KD. 2012. Shrub Willow. In: Kole C, Joshi CP, Shonnard DR, eds. *Handbook of Bioenergy Crop Plants*. Boca Raton, FL: CRC Press.

Soltis DE, Albert VA, Leebens-Mack J, Bell CD, Paterson AH, Zheng C, Sankoff D, dePamphilis CW, Wall PK, Soltis PS. 2009. Polyploidy and angiosperm diversification. *American Journal of Botany,* 96: 336–348.

Tamura S, Kudo G. 2000. Wind pollination and insect pollination of two temperate willow species, Salix miyabeana and Salix sachalinensis. *Plant Ecology,* 147: 185–192.

Volk TA, Abrahamson LP, Cameron KD, Castellano P, Corbin T, Fabio E, Johnson G, Kuzovkina-Eischen Y, Labrecque M, Miller R, Sidders D, Smart LB, Staver K, Stanosz GR, Rees Kv. 2011. Yields of willow biomass crops across a range of sites in North America. *Aspects of Applied Biology*: 67–74.

Zsuffa L, Mosseler A, Raj Y. 1984. *Prospects for interspecific hybridization in willow for biomass production.* Uppsala, Sweden: Swedish University of Agricultural Sciences, Department of Ecology and Environmental Research.

CHAPTER 4

JOINT LINKAGE AND ASSOCIATION MAPPING OF COMPLEX BIOMASS-RELATED TRAITS IN SHRUB WILLOW (*SALIX PURPUREA* L.)

Submitted for publication as: Carlson CH, Gouker FE, Crowell CR, Evans LM, DiFazio SP, Smart CD, Smart LB. 2018.

4.1 Abstract

The effects of global climate change will require development of more resilient plants that will only come from a better understanding of the genetic basis of complex adaptive traits related to biomass production. Increasing energy demands and the necessity to reduce greenhouse gas emissions are driving the development of lignocellulosic crops to provide an alternative to non-renewable energy sources. Joint linkage-association mapping holds immense potential for the evaluation of complex traits, because it exploits genetic variation garnered from both recent and past recombination events. Compared to conventional inbred crop plants, linkage disequilibrium is known to decay at much faster rates in undomesticated, outcrossing, and highly heterozygous species, such as willow (*Salix* spp.). Shrub willow is a sustainable and dedicated energy crop, bred to be fast-growing and high-yielding on marginal land without competing with food crops. Recent genomic advances have provided the biomass feedstock community with new tools to improve traits related to biomass yield and wood chemical composition. As biotic and abiotic stresses rise, rapid selection using an expanding suite of genomic tools will allow for the sustained improvement of non-model bioenergy crops in a rapidly changing climate. We utilized an annotated reference genome assembly of *Salix purpurea* L. for read mapping, variant discovery, and candidate gene identification among related and unrelated *S. purpurea* mapping populations. A panel of North American naturalized *S. purpurea* accessions and full-sib F_1 and

F_2 *S. purpurea* populations were extensively phenotyped for a suite of morphological, physiological, pest and disease resistance, and wood chemical composition traits collected from multi-environment and multi-year replicated field trials. Controlling for both population stratification and kinship in the association panel, a comprehensive mixed model analysis was used to dissect the complex genetic architecture and plasticity of these important traits. Individually, GWAS models tested differed in terms of power, but the combined approach, which corrects for yearly and environmental co-factors across datasets, improved the overall detection and resolution of associated loci, considering the small sample size in the association panel. Although there were few highly-significant GWAS loci located within support intervals of QTL for corresponding traits in the F_2, many large-effect QTL were identified for important biomass traits, as well as two major QTL hotspots on chr05 and chr10. This study provides the first comparison of linkage analysis and linkage disequilibrium mapping approaches in *Salix*, and highlights both the complementary and limits of these two approaches for elucidating the genetic architecture of complex bioenergy related traits of a woody perennial breeding program.

4.2 Introduction

Long-lived woody perennials such as trees and shrubs have proven to be reliable lignocellulosic feedstocks for second generation biofuel production (Cameron et al., 2008; Sannigrahi et al., 2010, Hanley and Karp, 2013). The use of biomass crops, such as *Salix*, under short rotation coppicing or short rotation forestry systems provides fast growth and high yields with relatively low agricultural inputs, which are characteristics that will help mitigate problems with increasing demand for food and energy, with decreasing availability of land and resources (Valentine et al., 2012). Significant effort is needed to develop accelerated crop breeding

strategies and with the advent of high-throughput, next-generation sequencing technologies (Goodwin et al., 2016), and low-cost genotyping protocols, the ability to accelerate breeding and selection, especially with non-model crops, will continue to improve (Kim et al., 2016).

Future breeding efforts will most likely follow statistically robust and computationally demanding methods for dissecting complex traits, such as genome-wide association studies (GWAS) (Soto-Cerda et al., 2014), linkage analyses (Grattapaglia and Sederoff, 1994), and genomic prediction methods (Meuwissen et al., 2001). Molecular breeding has already revolutionized genetic improvement programs by narrowing the gap between true biological regulation and model-based prediction of phenotypes. Marker-assisted selection (MAS) (Collard et al., 2008) at the seeding stage would be beneficial for perennial crops, especially for those with particularly long breeding cycles, e.g., *Salix*, *Populus*, *Eucalyptus*, and *Pinus* (Crossa and Federer 2012).

Current woody bioenergy crop improvement programs follow breeding and selection methodologies comparable with that of shrub willow (Karp et al., 2011). Traditionally, in shrub willow, this involves: (1) initial field-based selection of parent genotypes to generate full- or half-sib F_1 families, (2) growing F_1 progeny individuals in the greenhouse and transplanting to nursery beds or to pots in a pot yard, (3) selection of high-yielding and disease-resistant individuals from family-based field trial measurements, (4) upscaling selected individuals via clonal propagation for multi-environment and multi-year yield trials, and then (5) final selection of superior clones based on stability of yield. By selecting genotypes at the seeding stage, MAS can dramatically reduce the time and expense required of field selection trials (Allwright and Taylor, 2016).

Linkage-disequilibrium (LD)-based, association mapping exploits the natural genetic variation for a trait of interest by capturing historic recombination events among unrelated individuals, thereby offering a theoretically higher resolution to detect causal variants, compared with classic linkage analysis (LA), which relies on recent recombination events among related individuals (Pritchard et al., 2000, Mackay et al., 2009). Whole-genome association studies have the advantage of assaying the entire genome for trait-associated variants, so the number and specific choice of candidate genes is not restricted (Gaut and Long, 2003). In plant populations of unrelated individuals, and especially for highly-heterozygous obligate outcrossers, LD is expected to be low as a result of numerous historical recombination events (Khan and Korban, 2012). Relying heavily on the level of LD between a DNA marker and the causal variant, the success of LD mapping is directly related to genome-wide marker saturation. Yet, the increasing affordability of genome sequencing and ease of generating thousands of SNP-based markers in non-model crops provide great opportunity for discovery.

One benefit of LA is that it does not require a large number of markers, as does GWAS, yet mapping resolution can be low, because LA is limited by population size and the number of recombination events. With LD mapping, high-resolution maps can be obtained, particularly for long-living, perennial outcrossers, in which physical LD decays within one kilobase (Olson et al., 2010, Myles et al., 2011). Genome-wide association studies based on LD mapping therefore require a high marker density for detection of the causal variants, as realized by whole-genome resequencing in poplar (Evans et al., 2014). Studies combining LD and LA analyses have recently been carried out on several forest tree species, yielding promising results for growth traits in *Populus* hybrids (Du et al., 2016) and adaptive traits in *Picea mariana* (Prunier et al., 2013).

Association studies examining various growth, physiological, phenological, and wood composition traits have been conducted in a number of woody perennial species, such as *E. urophylla* (Denis et al., 2013), *P. balsamifera* L. (Olson et al., 2013), *P. tremula* (Ingvarsson, 2008), *P. trichocarpa* (Torr. & Gray) (Evans et al., 2014; McKown et al., 2014), and *P. deltoides* (Fahrenkrog et al., 2017), but so far only one association study has been conducted on willow, which examined *S. viminalis* (Hallingback et al., 2016). Residing in *Salix* sect. Vimen, *S. viminalis* is bred primarily for bioenergy and serves as the reference species for several European breeding programs (Karp et al., 2011). Trait-marker associations found in *S. viminalis* were related to various phenology and growth phenotypes, but the study only included 1,536 SNPs and used *Populus* as a reference for candidate gene selection, since a reference genome for *S. viminalis* was not then publicly available.

While there have been expression studies that have focused on assaying candidate genes involved in wood chemical composition (Serapiglia et al., 2012), as well as transcriptome-wide patterns of inheritance and allele-specific expression in F_1 and F_2 *S. purpurea* (Carlson et al., 2017), there have been no reports on large-scale, dense GWAS for biomass-related growth traits in outbred collections of the species. Here, we utilize a map-guided pseudomolecule assembly of the female *S. purpurea* genotype 94006 (Carlson et al., 2017; Zhou et al., 2018) for read mapping, variant detection, and candidate gene selection.

The objective of this study was to (1) identify QTL for a suite of morphological, phenological, physiological, and wood composition traits across different environments in *S. purpurea* using a full-sib F_2 bi-parental mapping population and a naturalized association panel, (2) discover significant SNPs and their associated gene candidates, and (3) compare and contrast the results from both mapping approaches for consistency and complementarity.

4.3 Materials and Methods

4.3.1 Plant Material and Growing Conditions

The field trials for the association panel were established using 20-cm cuttings of 112 accessions of *S. purpurea* that were planted at three experimental sites: Cornell AgriTech in Geneva, NY, Cornell University's Lake Erie Research and Extension Lab (CLEREL) in Portland, NY, and the West Virginia University Agronomy Farm in Morgantown, WV. Trials were planted in a randomized complete block design with six replicates of four-plant plots at each location in single-row spacing with 1.82 m between rows and 0.4 m between plants within rows. At the end of the establishment year, all plants were coppiced and trials were measured for a suite of biomass, architecture, phenology, physiology, composition, and pathology traits (Table 4.1) using the inner two plants of each four-plant plot across all sites in 2013 and 2014, and then mechanically harvested and weighed in 2015. Prior to re-growth of the second rotation in 2015, 112 kg ha^{-1} N-P-K fertilizer was applied to half of the replicates at each location to test for nitrogen utilization.

A full-sib F_1 family (Family 82) was generated from a cross between female *S. purpurea* 94006 and male *S. purpurea* 94001, both collected from naturalized *S. purpurea* in upstate NY. Two F_1 offspring from this cross, female *S. purpurea* 'Wolcott' (clone 9882-41) and male *S. purpurea* 'Fish Creek' (clone 9882-34), were crossed to generate a full-sib, intraspecific F_2 *S. purpurea* family (Family 317). All progeny individuals and their parents were planted in nursery beds at Cornell AgriTech, Geneva, NY. In 2014, dormant whips from 497 F_2 *S. purpurea* progeny individuals and the parents and grand-parents of the F_2 pedigree were collected from their nursery beds, and stored at -4°C prior to spring field planting.

Table 4.1 Traits collected and their abbreviations and units.

Trait	Abbreviation	Units
Biomass		
Plot height	HT	m
Stem diameter	DIA	mm
Stem number	STNo	#
Stem area	SA	cm^2
Stem volume	VOL	cm^3
Stem internode length	INLEN	cm
Plot yield	YLD	dry Mg ha^{-1}
Foliar		
Leaf area	LFA	cm^2
Leaf length	LFL	cm
Leaf width	LFW	cm
Leaf perimeter	LFP	cm
Leaf shape factor	LFF	
Leaf aspect ratio	LFR	
Leaf dry weight	LFDW	g
Specific leaf area	SLA	$cm^2\ g^{-1}$
Architecture		
Crown diameter	CDIA	cm
Crown form	FORM	degrees °
Chemical Composition		
Hemicellulose	HCL	%
Cellulose	CLS	%
Lignin	LIG	%
Ash	ASH	%
Wood density	DEN	g cm^{-3}
Phenology		
Vegetative phenology	VPHE	date
Floral phenology	FPHE	date
Individual sex	SEX	(F, M, H)
Physiology		
SPAD (August)	SPAD-1	SPAD units
SPAD (September)	SPAD-2	SPAD units
Stomatal conductance	COND	mmol $m^{-2}\ s^{-1}$
Stem color	STC	(0, 1, 2)
Pathology		
Rust severity	RUST	%

A field trial with the F_2 progeny, parents and grand-parents was established at Cornell AgriTech in Geneva, NY in a randomized complete block design with four replicate blocks of three-plant plots. To avoid edge effects, *S. purpurea* genotypes 'Fish Creek' and 94006 were planted as border rows along the east and west sides of the trial, respectively, and the north and south ends were buffered by a single row of genotype 94006. Within-row spacing was 0.4 m and spacing between rows was 1.82 m. The soil at the field site is Odessa silt loam with a depth to water table of 25 to 45 cm. For additional site characteristics, see Serapiglia et al. (2014).

4.3.2 DNA Isolation and Sequencing

Briefly, tissue for DNA extraction was collected from young leaves and shoot tips, flash frozen in liquid nitrogen, then ground to a fine powder with a Geno/Grinder® (SPEX SamplePrep, Metuchen, NJ, USA), and genomic DNA extracted using the DNeasy® Plant Mini Kit (QIAGEN Inc., Valencia, CA, USA). The quality of DNA was checked by agarose gel electrophoresis and quantity was estimated using a NanoDrop ND-1000 spectrophotometer (Thermo Scientific; Wilmington, DE, USA). Library and sequencing preparation was based on a 48-plex (association panel) or 96-plex (F_2 family 317) genotyping-by-sequencing (GBS) protocol according to Elshire et al. (2011). *ApeK*I served as the restriction enzyme for the association panel, whereas both *ApeK*I and *Eco*T22I restriction enzymes were used for the F_2 family. Resulting libraries were sequenced on the Illumina HiSeq 2000 (Illumina, Inc.; San Diego, CA, USA) platform at the Cornell University Biotechnology Resource Center (Ithaca, NY, USA).

4.3.3 Read Mapping and Variant Discovery

Variant discovery and filtering was performed with the TASSEL v3.0 GBS Discovery Pipeline (Bradbury et al., 2007), along with custom Perl and R scripts (available online at:

https://github.com/Willowpedia). Raw reads from FASTQ files were trimmed to 64 bp and were processed to create a set of unique sequence tags (min.cov = 5, n = 4,550,690). Genotypes were called via physical alignment to the female *S. purpurea* 94006 reference genome assembly with BWA *mem* (Li and Durbin, 2009). The physical positions of UNEAK tags were produced by a local BLASTN query of *ApeK*I and *Eco*T22I tags to the genome reference. For the association panel, SNPs were retained in individuals with a call rate of < 90%, minor allele frequency (MAF) < 0.05, and maximum proportion of 50% missing data, which provided a set of 103,180 high-quality SNPs. Marker imputation was conducted using a LD-kNNi approach (kNN = 5, LD sites = 20, LD window = 10 Mb), outlined in Money et al. (2015). Imputation accuracy was assessed by masking high-confidence variants (mapq > 30, cov > 10) on the original dataset in TASSEL v5 to obtain accuracy metrics. For the F_2 family, reference and non-reference (UNEAK) mapping approaches produced ~300,000 SNPs and ~12,000 SNPs, respectively. Overall, the enzyme *Eco*T22I gave better mean depth of coverage (~11×) but fewer SNPs, whereas *ApeK*I gave nearly five-times as many SNPs as *Eco*T22I, and had less missing data overall.

4.3.4 Linkage Disequilibrium

To evaluate the marker resolution expected during GWAS, LD (r^2) was calculated for all pairs of SNPs after imputation within 5 kb using PopLDdecay v3.3 (https://github.com/BGI-shenzhen/PopLDdecay). Only markers with minor allele frequency values above 0.05 and having less than 25% missing data among the sample set were included for LD analyses. The maximum value of r^2 was calculated based on SNP pairs within 1 kb and a LD decay curve was plotted based on r^2 and the distance between pairs of SNPs and a non-linear regression curve was fitted.

4.3.5 Linkage Map Construction

Marker coding was based on multiple GBS runs of the reference grandparents and parents of the F_2 family. Only biallelic and unambiguous sites in both grandparents were considered. Any markers deviating from Hardy-Weinberg equilibrium ($P < 1\times10^{-3}$) and individuals with excessive missing data or skewed genotype proportions were removed prior to map construction in R/qtl (Broman et al., 2003). For the reason that GBS has a general tendency to under-call heterozygous sites in outcrossing species, like willow, GBS markers were minimally imputed one chromosome at a time with *imputeByFlanks* and error-corrected with *correctUnderCalledHets* and *correctStretches* (maxHapLength = 3) functions in the ABHgenotypeR, as described in Furuta et al. (2017).

Markers were partitioned and ordered in linkage groups using a minimum spanning tree approach with MSTmap (Wu et al., 2008b) in ASMap (Taylor and Butler, 2017), with the parameters: pop.type = RIL2, dist.fun = kosambi, *p*-value = 1×10^{-12}, miss.thresh = 0.15, noMap.dist = 15, noMap.size = 5, and detectBadData = true. Final intercross genotype percentages were 25.7%, 48.7%, and 25.6%, for AA (94006/94006), AB (94006/94001), and BB (94001/94001), respectively. The final linkage map consisted of 6,045 markers and 26 linkage groups, representing all 19 haploid chromosomes of *S. purpurea* (Figure 4.1). Linkage groups ranged from 11 cM (LG 7B) to 369 cM (LG 2), with a total map length of 3465 cM and an average marker density of 1.7 cM (Table 4.2). A number of GBS markers incorporated into the master linkage map mapped to a total of 155 unplaced physical scaffolds in the *S. purpurea* v1.0 reference genome assembly.

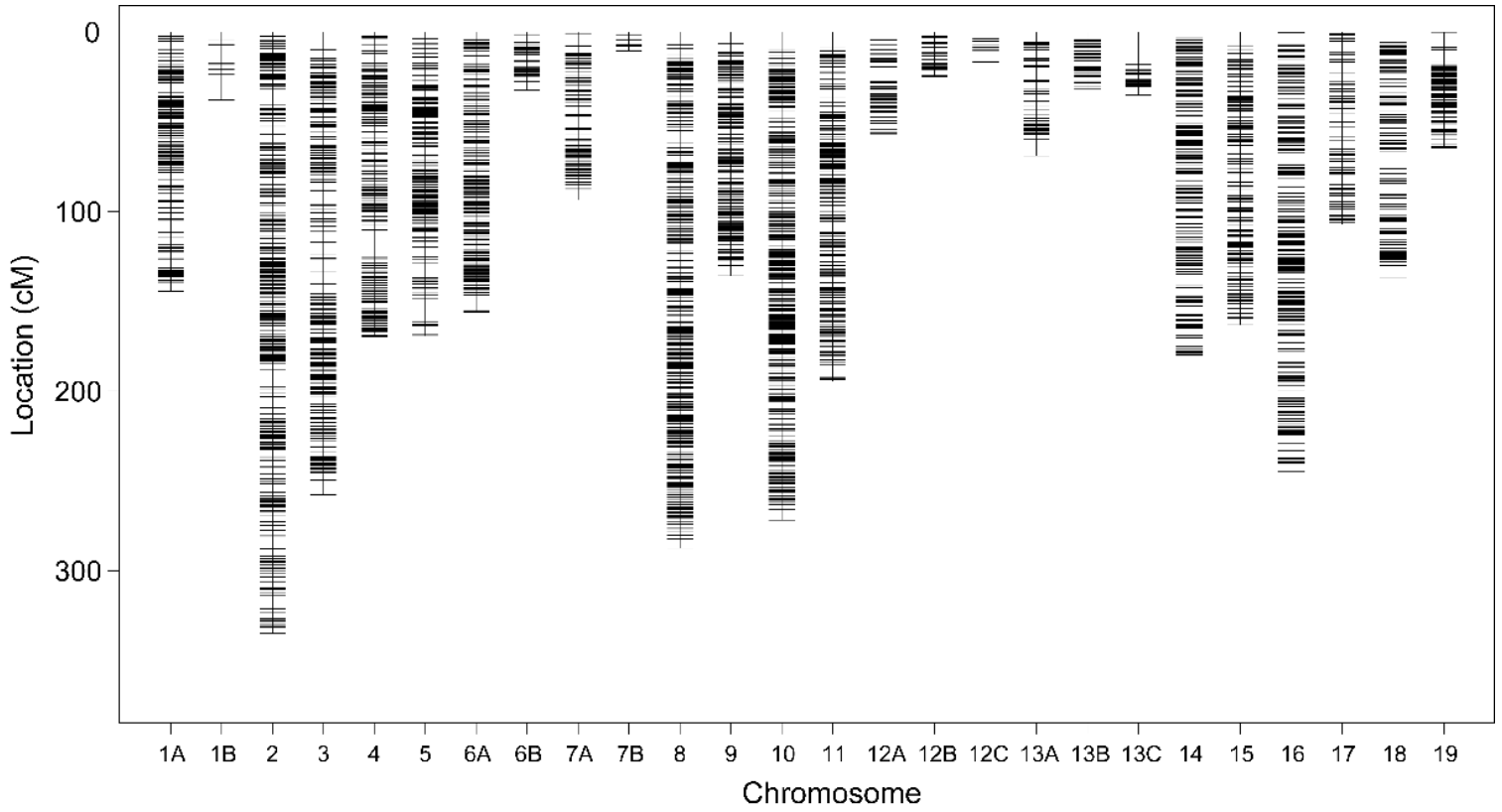

Figure 4.1 Linkage groups of the F_2 genetic linkage map representing the 19 chromosomes of *S. purpurea* and are named according to respective physical chromosomes.

Table 4.2 Linkage map statistics by linkage group.

Chromosome	Linkage Group	Markers	Distance (cM)	Markers / cM
1	1A	204	144.3	1.42
1	1B	9	37.90	0.24
2	2	573	369.9	1.55
3	3	397	257.8	1.55
4	4	295	169.6	1.74
5	5	280	169.2	1.66
6	6A	376	155.9	2.42
6	6B	61	32.50	1.88
7	7A	140	93.50	1.50
7	7B	7	10.70	0.66
8	8	603	287.4	2.10
9	9	296	135.8	2.19
10	10	620	271.9	2.29
11	11	309	194.6	1.59
12	12A	66	57.20	1.16
12	12B	40	25.10	1.60
12	12C	17	16.70	1.02
13	13A	73	69.00	1.06
13	13B	53	31.90	1.66
13	13C	56	35.60	1.58
14	14	312	182.9	1.71
15	15	308	163.1	1.89
16	16	464	244.9	1.90
17	17	133	107.2	1.25
18	18	206	136.7	1.51
19	19	148	64.70	2.29

4.3.6 Phenotyping

During the dormant period after each growing season, diameters (DIA, cm) of stems ≥5 mm were measured at 30 cm from the base of the plant using Masser Racal 500 digital calipers and stem number was counted for each plant (Masser, Rovaniemi, Finland). Total stem area (SA, cm^2) per plant was also calculated using the stem diameter values. Maximum stem height (HT, m) of every plot was recorded using a measuring rod (Crain Enterprises, Inc., Mound City, IL). In July of each year, internode length (INLEN, cm) was measured within the middle third of the tallest stem of each plot and the length of five internodes were recorded. Accounting for different phyllotactic patterns, alternate leaves were counted using five alternate buds or leaves from the first designated bud/leaf, whereas opposite leaves or buds were counted as one node.

Yield of each plot in the three trials containing the diverse *S. purpurea* collection was measured after the second year of post-coppice by harvesting and weighing all four plants in each plot using the Ny Vraa JF192 harvester (Ny Vraa Bioenergy, Tylstrup, Denmark). Chips were collected in a plastic bin mounted on Avery Weigh-Tronix weigh cells (Fairmont, MN), and the total wet weight of the chip biomass of each plot was recorded. A sub-sample of fresh chip biomass (~1 kg) was collected for each plot, weighed after harvest, oven-dried at 65ºC to a constant weight, and dry weight recorded to determine moisture content at harvest. The moisture content was then used to estimate plot dry weights from the measured fresh weights. For all plots, dry biomass yield was calculated and expressed in dry Mg ha^{-1} based on plot area.

At the end of the second growing season, crown diameter (CDIA, cm) was measured using modified Haglöf Mantax forestry calipers (Haglöf Sweden AB, Långsele, Sweden). Stool diameters were measured at 15.24 cm (6 in) above the soil, which is the average height of a shrub willow harvester. Crown form (FORM, degrees °) was calculated by multiplying the

arctangent2 of one-half CDIA and the fixed distance at which CDIA was measured (15.24 cm) by 180/π, to obtain the angle of the stem branching relative to the soil.

Leaf area (LFA, cm^2), length (LFL, cm), width (LFW, cm), and perimeter (LFP, cm) were measured on mature leaves at mid-canopy level on the tallest stem of each plant per plot using a CID CI-203 laser leaf area meter (CID Bio-Science, Inc., USA). The same measurement leaves were collected, dried at 65°C, and weighed. Leaf area and dry weight (LFDW, g) were used to calculate specific leaf area (SLA) (cm^2 g^{-1} dry wt). Leaf aspect ratio (LFR) is the ratio of LFL to LFW, and leaf shape factor (LFF) is the ratio of LFA to LFP, but corrected, $4\pi[LFA/LFP^2]$, so that the LFF of a circle is equal to one.

Physical and chemical wood properties were measured for four replicates in each of the three trials with the diverse *S. purpurea* collection. Stem segment samples were collected in the dormant period after each growing season using sampling methods previously described (Liu et al., 2015) and were stored frozen at -4°C until they were processed. The specific gravity of each sample was measured by volumetric displacement (TST om-06, 2006). In 2014, a modified method of measuring specific gravity was used where the volume of water displaced was weighed for added precision. Following specific gravity determination, stem segments were oven-dried at 65°C to a constant weight and then rough milled to a 5 mm particle size with a Retch SM300 cutting mill (Retch, Haa, Germany) and were further comminuted to <0.5 mm particle size by fine milling with the IKA MF 10.1 knife mill (IKA, Wilmington, NC) for compositional analysis. Approximately 20 mg of each milled stem sample was analyzed with a Thermogravimetric Analyzer (TGA) Q500 instrument and Universal Analysis 2000 version 4.5A software (TA Instruments, New Castle, DE), as previously described (Serapiglia et al., 2009).

Hemicellulose, cellulose, lignin, and ash content were determined as a percentage of total dry biomass for each sample, as previously described in Serapiglia et al. (2014).

Stomatal conductance (COND, mmol m^{-2} s^{-1}) was measured on the abaxial side of the leaf with a leaf porometer (SC-1 Leaf Porometer, Decagon, Pullman, WA) on the uppermost fully expanded leaf of the tallest stem of the plant. A non-destructive proxy for leaf nitrogen status was measured with a portable chlorophyll meter (SPAD-502, Minolta Osaka Co., Ltd., Japan) where readings were collected from three leaves along the length of the tallest stem from the upper, middle, and lower canopy levels and averaged for each plot. Canopy color (RGB-15) in the trial with the of F_2 population was determined by plot using aerial images collected with a gimbal-mounted 14 M F/2.8 140° FOV camera (w/ lens stabilization) on a Phantom 2 Vision+ (DJI, Nanshan District, Shenzhen, China) quadcopter. To account for any variation, three replicate images were taken for each interval at a fixed altitude (37 m) along the length of the field trial (365 m) in late-July 2015. An overlap of each interval was required to properly interleave into a single image. Images were lens corrected using the DJI Vision plugin, ordered, and interleaved using Adobe Photoshop CS6 (Adobe Systems Incorporated, San Jose, CA). The resulting interleaved full-field images were converted into separate RGB channels and analyzed by plot using a colorimetric scale based on green pixel density in ImageJ v1.47 (Rasband, 1997-2016; Schneider *et al.*, 2012). Excluding aisles and border plants, a coordinate grid of the field was used to calculate average pixel density for each plot. Stem color (STC-0) was scored on establishment-year growth using a qualitative scale (0 = entirely green, 1 = intermediate, 2 = entirely red).

Floral and vegetative bud break were observed and scored using a 0-5 scale only in the second year of growth due to the sparsity of floral buds in the first year. The established scale

used for phenology ratings was modified from Saska and Kuzovkina (2010). Both floral and vegetative phenology was surveyed once a week for five weeks and was recorded as the day of the year for a given rating that was observed. All observations occurred until all stage 5 scores were recorded for every genotype. For all trials, the sex of each genotype was recorded. While the floral phenology and morphology of *S. purpurea* catkins are reliably sexually dimorphic, we identified three accessions in the association panel (94003, 00-22-002, and 06-01-003) which routinely displayed mixed proportions of female and male flowers along identical catkin inflorescences. Perfect flowers were not observed, rather, male and female flowers were mixed within catkins. The proportions often varied among catkins, which were primarily female at the proximal base and male at the distal tip. SNP analysis of these three accessions determined that they represented a single clone, 94003, which is a hermaphrodite.

In September 2015, willow leaf rust (*Melampsora* spp.) was visually scored in two (Geneva and WVU) of the three association trials by Chris Smart, whereas all three trials were scored in September 2017 by Chase Crowell. Willow leaf rust was scored for both 2015 and 2017 for the F_2 trial by Chris Smart and Chase Crowell, respectively. Percent rust severity (RUST, %) was visually scored for each plot based on total leaf area infected. In 2017, maximum RUST scores in the association population were capped at 50% due to disease-related defoliation of heavily infected shrubs. Disease ratings were completed within a biologically relevant time period within and among field trials.

4.3.7 Statistical Models and Analysis

All statistical analyses were conducted within the open-source statistical computing environment, R (R Core Team, 2015).

Variance components were estimated using REML with the *lmer* function in the R package, lme4 (Bates et al., 2015), using the linear mixed model:

$$\boldsymbol{y_{ijk} = \mu + g_i + \rho_j + \rho(\gamma)_{jk} + \varepsilon_{ijk}}$$

where $\boldsymbol{\mu}$ is the population mean, $\boldsymbol{g_i}$ is the effect of genotype, $\boldsymbol{\rho_j}$ is the effect of block, $\boldsymbol{\rho(\gamma)_{jk}}$ is the effect of block within environment, and $\boldsymbol{\varepsilon_{ij}}$ is the random error, which are assumed to be independent and identically distributed.

Following the approach outlined in Velazco et al. (2017), spatial trends (row and column) in the F_2 field trial were modeled as two-dimensional Penalized (P)-splines, using *SpATS* and *SAP* functions (n.seg = (16, 64), tolerance = 1×10^{-6}) in the SpATS package (Rodríguez-Álvarez et al., 2015, Rodríguez-Álvarez et al., 2017).

Prior to genetic mapping, a Shapiro-Wilks (Shapiro and Wilk, 1965) test for normality was performed on each quantitative trait (BLUEs or BLUPs of genotypes from the above model) with *shapiro.test* in R. Trait distributions showing a significant departure from normality ($P < 0.05$) were Box-Cox transformed (Box and Cox, 1964) with appropriate transformation parameters, lambda (λ) and gamma (γ), estimated with the *powerTransform* function in the car package (Fox and Weisberg, 2011). For positive responses, the 'bcPower' method was used, and for those including negative responses, the two-parameter 'bcnPower' method was used. Clonality in the association panel was assessed by calculating identity by state (IBS) between each pair of individuals. No efforts were made to merge markers for clonal entries, rather, for those exceeding a threshold of ≥ 0.96, the individual with the least missing marker data was retained and data from other accessions of each clone was not considered.

Subsequent genotype best linear unbiased predictors (BLUP), from the above model were used in GWAS of 78 clones after ramets were removed and calculated in lme4 using the *lmer*

function (Bates et al., 2015). In order to control confounding effects and improve statistical power while reducing the incidence of inflated *P*-values, GWAS was conducted using the model selection algorithm, the Fixed and random model Circulating Probability Unification (FarmCPU) (Liu et al, 2016) which takes into account the confounding problem between covariates and test marker by using both Fixed Effect Model (FEM) and a Random Effect Model (REM). Additionally, the first three principal components calculated using GAPIT (Lipka et al., 2012) were used as covariates to control for population structure. Three models were tested concurrently in the FarmCPU package, GLM, MLM and FarmCPU. The default *p*-value threshold set in FarmCPU uses a Bonferroni-corrected threshold ($\alpha = 0.01$). However, the Bonferroni-corrected multiple testing threshold is overly strict when the LD among genotypic markers is large, so the threshold was calculated using 1,000 permutations (p.threshold = 0.05/number of markers). The threshold calculated by FarmCPU for the given traits, $-\log_{10}$(*p*-value) = 6.31, was used as a cut-off to determine multiple-trait associations. To determine which models and corrected parameters best fit the data, observed and expected $-\log_{10}$(*p*-value) distributions for each SNP association were plotted as quantile-quantile (QQ)-plots. In order to account for sex-specific associations, individual sex was used as a covariate in all traits assayed.

Linkage analysis in the F_2 family was performed entirely in R/qtl (Broman et al., 2003, Broman, 2018). Genotype probabilities were calculated using *calcgenoprob* (step = 1.5; off.end = 0; stepwidth = fixed; map.function = kosambi). Single QTL models were run using the *scanone* (method = EM; n.perm = 1000; max.it = 10000; tol = 1×10^{-6}) function, then refined using *makeqtl* and *refineqtl*. In addition, the leave-one-chomosome-out (LOCO) method was compared to single scan results, which utilizes marker-based kinship in the model and excludes the chromosome in which the top LOD marker was located, reusing variance parameters fitted

once per left-out chromosome. Multiple interactive QTL models were fit using *makeqtl* and *fitqtl*. Percent phenotypic variance (% Vp) from individual and full QTL models are reported from *fitqtl* results. Both *bayesint* (prob = 0.95) and *lodint* (drop = 1.5) functions were used to calculate LOD support intervals from the output of *refineqtl* for each chromosome exceeding permutation thresholds ($\alpha = 0.05$), and ranged from 4.1 to 4.4.

If a trait significantly differed by sex (Wilcoxon $P < 0.05$), individual sex was added to each model as a covariate to avoid confounding effect of sex-linkage. Traits identified as sexually dimorphic were: FPHE-2, HCL-2, LIG-2, CDIA-1, FORM-1, LFDW-1, LFF-2, SLA-1, and SLA-2, as well as more weakly-associated ($P < 0.1$) traits CLS-2, CDIA-2, HT-0, and HT-1.1.

4.4 Results

4.4.1 GWAS Results

Of the 112 accessions planted in the association trial, an analysis of identity by state (IBS) revealed eight instances of clonality with a threshold of ≥ 0.96 resulting in the removal of 33 accessions from the analysis. One male clone was represented 16 times in the trial. Clonality can be explained by vegetative propagation across the landscape, most likely conducted by humans, since multiple accessions were collected from natural sites at great distances from each other. For others that were obtained from collaborators or nurseries, clonality was likely due to propagation and renaming of other accessions in the collection. This resulted in a final tally of 79 unique genotypes that were subjected to GWAS.

The dataset obtained using the reference based GBS filtering pipeline yielded 103,180 SNP markers with an average nucleotide diversity $\pi = 0.32$. Filtering criteria were selected to remove markers with MAF < 0.05, and the MAF distribution of the remaining marker at MAF <

0.10 was 18.4% and the MAF > 0.25 was 37.4% with the average MAF of 0.22. The heterozygosity rates for each genotype ranged from 0.17 to 0.49 with an average heterozygosity of 0.27. The average density of SNPs corresponded to 1 marker every 3.8 kb. Linkage disequilibrium analysis showed that 19.2% of the marker pairs were in LD with a majority of markers exhibiting average r^2 values of 0.24.

GWAS had limited power to detect significant associations of small or moderate effect for a suite of traits (Figure 4.2, Table 4.3), due to the small sample size of this population (n = 79). In order to avoid detection of false positives, a series of GLMs and MLMs were used to increase statistical power and correct for kinship and population structure. All models for each trait were evaluated based on the fit of the model to the data and resulted in the 'FarmCPU' method performing well overall with an average genomic inflation factor of $\lambda_{gc} = 1.11$.

Most of the marker associations that reached genome-wide significance were located within genic regions (5'UTR, CDS, intron, or 3'UTR) and candidate genes were inferred based on functional annotation of the *S. purpurea* genome. Overall, 99 significant associations ($p < 5\times10^{-7}$) were detected on 15 of the 19 *Salix* chromosomes, with no associations mapping to chromosomes 9, 14 and 18. Of all significant markers associated with 19 of the measured traits, only six of the significant SNP associations did not intersect or were distant from gene models in the reference annotation. There were 82 SNPs that were within genic regions and the remaining SNPs were positioned between genes. For SNPs within genes, there were 60 unique genes associated with > 50% of the traits assayed.

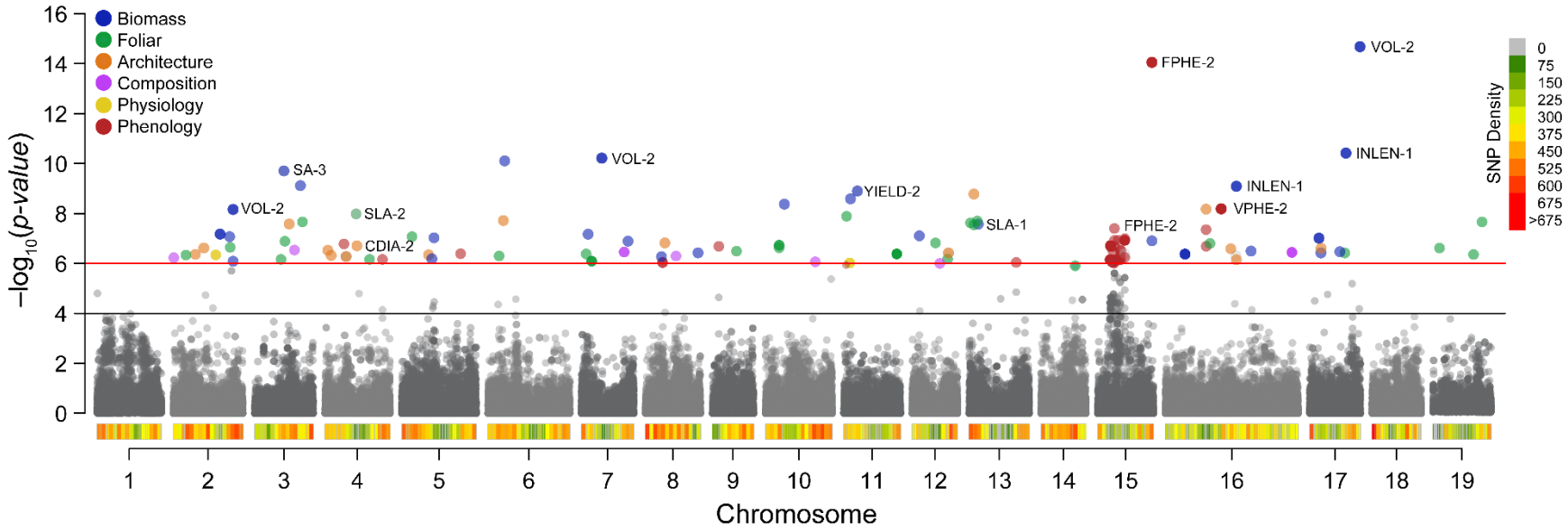

Figure 4.2 Manhattan plot of GWAS hits by trait class in the association panel. Points signify significant SNP-trait associations ($P < 5\times10^{-7}$) using the FarmCPU model within each class and are colored by the corresponding trait (see legend). To improve the visual representation of GWAS hits, nearby SNPs < 100 Kb were concatenated into a single representative location. SNP densities are illustrated above chromosomes labels as heatmaps, which are binned according to the legend to the right of the Manhattan plot.

Table 4.3 Significant SNP-trait associations from the *S. purpurea* association panel.

Trait	Year	Site	Chrom	Pos (Mb)	Effect	$-\log_{10}(p)$
Biomass						
HT	1	PTL	15	13.21	-0.33	6.9
INLEN	1	GVA	2	5.4	-2.97	6.4
INLEN	1	GVA	2	7.46	2.1	6.6
INLEN	1	GVA	4	1.51	3	6.3
INLEN	1	GVA	4	7.79	2.6	6.7
INLEN	1	PTL	3	8.41	-1.69	7.6
INLEN	1	PTL	5	6.51	-0.77	6.3
INLEN	1	PTL	8	4.72	-0.85	6.8
INLEN	1	PTL	16	13.64	-1.08	8.2
INLEN	1	PTL	16	17.33	1.17	9.1
INLEN	1	PTL	17	2.71	-0.75	6.6
INLEN	1	PTL	17	8.78	-0.98	10.4
SA	2	PTL	2	11.43	-3.91	9.1
SA	2	PTL	2	13.75	-1.68	7.1
SA	2	PTL	10	4.58	2.46	8.4
SA	2	PTL	15	13.21	-4.59	14.1
SA	2	WVU	16	4.82	6.58	6.4
SA	2	WVU	16	4.82	6.58	6.4
SA	2	WVU	16	4.82	6.58	6.4
SA	2	WVU	17	2.21	-7.21	7
SA	2	WVU	17	2.21	-7.21	7
SA	3	GVA	3	7.12	-1.16	9.7
SA	3	GVA	7	11.34	0.81	7.8
SA	3	GVA	13	2.22	-1.16	7.6
SA	3	GVA	16	20.91	0.7	6.5
DIA	1	WVU	8	12.81	21.43	6.7

Table 4.3
(continued)

Trait	Year	Site	Chrom	Pos (Mb)	Effect	$-\log_{10}(p)$
DIA	2	PTL	2	11.43	-55.07	6.8
DIA	2	PTL	2	11.43	55.07	6.8
DIA	3	GVA	7	11.34	26.18	6.5
DIA	3	GVA	16	20.91	27.31	6.6
STNo	1	WVU	8	12.81	2.54	6.6
STNo	3	WVU	7	11.34	1.96	6.6
VOL	1	PTL	2	11.43	-249.3	6.3
VOL	1	PTL	2	11.43	249.3	6.3
VOL	1	PTL	17	7.3	242.8	6.5
VOL	1	WVU	2	14.6	118.4	6.5
VOL	2	GVA	5	7.33	-806.9	6.2
VOL	2	GVA	17	2.68	1196.7	6.4
VOL	2	PTL	3	11.14	-261.4	9.1
VOL	2	PTL	7	1.61	190.6	7.2
VOL	2	PTL	7	4.94	269.9	10.2
VOL	2	PTL	11	1.6	336.2	8.6
VOL	2	PTL	11	12.28	-264.3	8.8
VOL	2	PTL	12	1.73	199.6	7.1
VOL	2	PTL	17	7.3	611	14.7
VOL	2	WVU	2	14.6	351.9	8.2
VOL	2	WVU	5	7.84	599.1	7
VOL	2	WVU	6	4.07	401.5	10.1
VOL	2	WVU	17	9.85	218.2	6.6
VOL	3	GVA	7	11.34	188.1	6.7
YIELD	2	GVA	2	11.43	-1.48	6.9
YIELD	2	GVA	2	11.43	-1.48	6.9

Table 4.3
(continued)

Trait	Year	Site	Chrom	Pos (Mb)	Effect	$-\log_{10}(p)$
YIELD	2	GVA	2	11.43	1.48	6.9
YIELD	2	GVA	2	11.43	1.48	6.9
YIELD	2	PTL	2	11.43	-0.6	7.3
YIELD	2	PTL	11	3.3	0.41	8.9
YIELD	2	PTL	15	13.21	-0.39	6.5
Foliar						
LFDW	1	GVA	17	8.52	0.02	6.4
LFF	1	PTL	15	3.3	-0.02	6.1
LFF	1	PTL	16	10.9	-0.03	6.8
LFL	1	GVA	2	13.9	1.38	6.3
LFR	1	GVA	3	7.38	-1.35	6.9
LFR	1	GVA	19	1.6	-0.48	6.6
LFR	2	GVA	7	1.12	-1.27	6.4
LFR	2	GVA	11	12.89	-1.27	6.4
LFR	2	GVA	11	12.89	-1.27	6.4
LFR	2	GVA	11	12.89	-1.27	6.4
LFR	2	GVA	11	12.89	1.27	6.4
LFR	2	GVA	11	12.89	1.27	6.4
LFW	1	PTL	2	3.09	-0.1	6.3
SLA	1	PTL	3	11.63	3.7	7.7
SLA	1	PTL	5	2.48	2.17	7.1
SLA	1	PTL	7	2.48	3.26	8.7
SLA	1	PTL	13	1.15	4.86	8.8
SLA	1	PTL	19	11.97	-2.89	7.7
SLA	2	PTL	10	3.31	3.5	6.7
SLA	2	PTL	10	3.31	-3.46	6.6

Table 4.3
(continued)

Trait	Year	Site	Chrom	Pos (Mb)	Effect	$-\log_{10}(p)$
SLA	2	PTL	10	3.31	3.5	6.7
SLA	2	WVU	4	7.61	-2.11	8
SLA	2	WVU	11	0.65	-2.69	7.9
SLA	2	WVU	13	0.33	-2.9	7.6
Architecture						
CDIA	2	PTL	4	0.7	5.16	6.6
CDIA	2	WVU	4	5.15	5.5	6.4
CDIA	2	WVU	4	5.15	-5.5	6.4
CDIA	2	WVU	4	5.15	5.5	6.4
CDIA	2	WVU	4	5.15	-5.5	6.4
Composition						
ASH	1	WVU	16	30.88	0.25	6.5
ASH	1	WVU	16	30.88	-0.25	6.5
ASH	1	WVU	16	30.88	-0.25	6.5
LIG	2	WVU	3	9.7	-0.57	6.5
Phenology						
FPHE	2	GVA	15	4.07	-1.65	18
FPHE	2	WVU	15	6.59	7.76	6.4
VPHE	2	WVU	4	4.63	-0.87	6.8
VPHE	2	WVU	5	14.32	0.75	6.4
VPHE	2	WVU	16	9.95	1.22	8.2
SPAD1	1	PTL	10	15.97	-0.41	5.4
SPAD1	1	PTL	17	10.29	-0.62	5.2
SPAD2	2	PTL	2	10.36	-1.98	6.3
SPAD2	2	PTL	2	10.36	-1.98	6.3

A majority of these SNPs were associated with biomass-morphology traits, but also significant trait associations were found with SPAD, ASH, LIG, FPHE and VPHE. In the association panel, sex dimorphism for phenological traits, FPHE-2 and VPHE-2, were more pronounced in the WVU trial, compared to the Geneva trial, with differences in means (female − male, $P < 0.01$) ranging from 5.1 to 11.7 and 1.4 to 1.8, respectively (Table 4.4). In addition, the difference in the number of days of FPHE-2 and VPHE-2 ranged from −10 to −3.7, respectively. All significant associations with crown diameter occurred on chr04 as well as SNPs associated with SLA and INLEN. Other growth traits including SA, SDIA, STNo, and VOL had significant associations across 10 chromosomes, but with overlapping genomic regions for these traits with several of the SNPs falling within similar candidate genes such as DNA binding domains and raffinose synthase genes. Significant SNPs were found with associations for yield across two of the experimental sites on chromosomes 2, 11, and 15 which were associated with candidate genes related to protein kinases and protein binding domains.

Table 4.4 Phenological differences by sex (female − male) in the *S. purpurea* association panel.

Trait	**Location**	**Estimate**	**± SE**	**CI**	***t*-value**	***p*-value**
FPHE	Geneva	5.1	0.52	4.08 - 6.16	9.79	< 0.001
	WVU	11.7	2.04	7.63 - 15.8	5.73	< 0.001
VPHE	Geneva	1.4	0.59	0.23 - 2.61	2.38	0.02
	WVU	1.8	0.59	0.65 - 3.03	3.09	0.003
VPHE - FPHE	Geneva	-3.7	0.69	-5.09 - -2.33	-5.35	< 0.001
	WVU	-9.9	1.86	-13.6 - -6.17	-5.32	< 0.001

4.4.2 Linkage Mapping Results

Biomass QTL were well-represented on linkage groups representing *Salix* physical chromosomes 4, 5, 6, and 10 (Figure 4.3, Table 4.5).

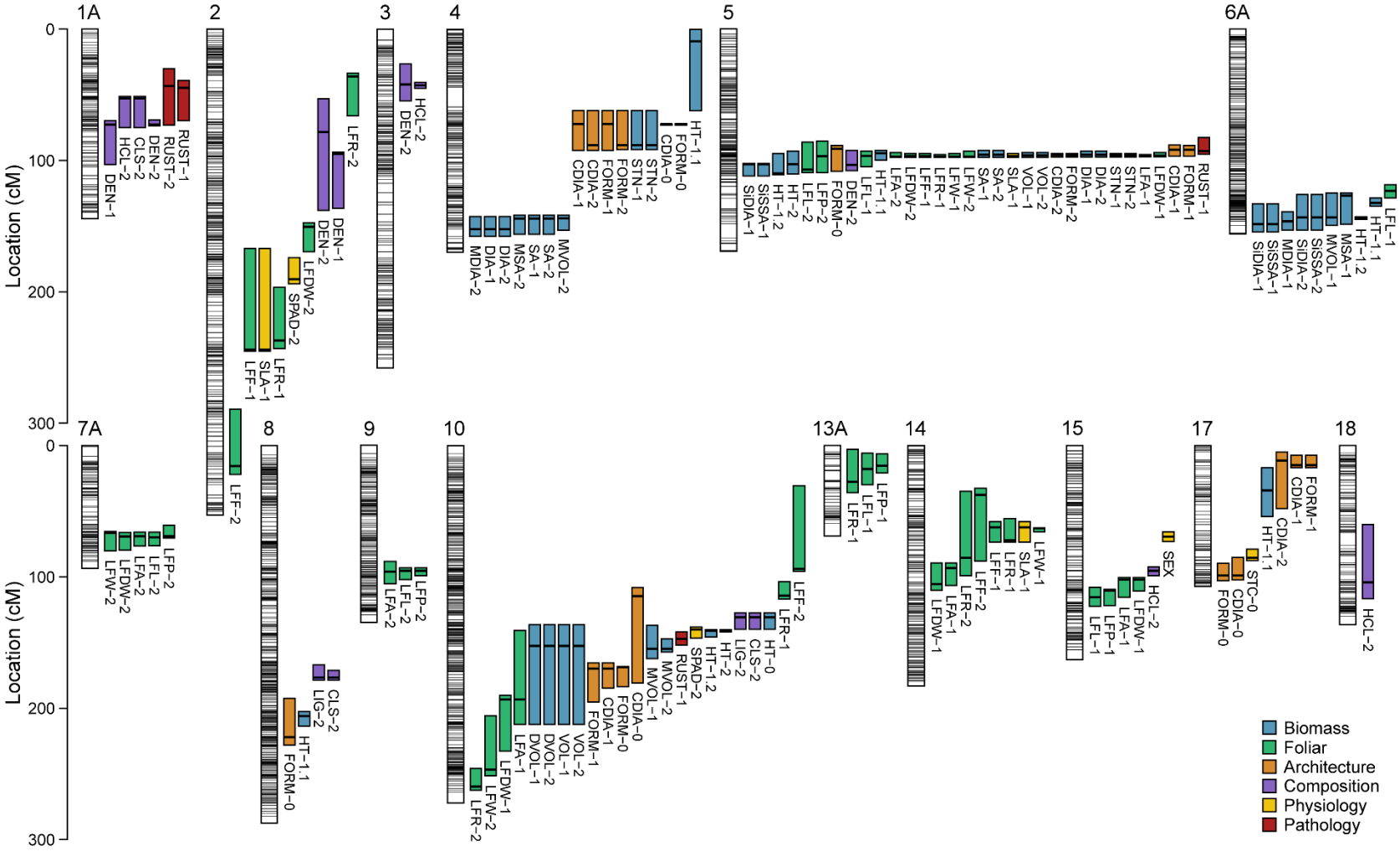

Figure 4.3 LOD support intervals of QTL for biomass-related traits anchored to the F_2 *S. purpurea* linkage map. Linkage groups are labeled according to their respective physical chromosome. Bars to the right of each linkage group represent LOD support intervals for respective trait QTL. Only QTL for a subset of traits with significant LOD support (LOD > 4.1) are shown. Horizontal lines within support intervals indicate the peak LOD position. Support interval bars are colored according to their phenotypic class (see legend).

Table 4.5 Mapped QTL for biomass-related traits for the F_2 *S. purpurea* mapping population.

Trait Class	Trait-Year	LG	Pos	CI	LOD	mu	a	d	h^2	Marker Vp (%)	Full Vp (%)
Architecture	CDIA-0	4	72.7	72-73	9.8	-4.6E-03	5.4E-02	1.7E-02	0.56	8.67	38.55
Architecture	CDIA-0	5	96	91-109.6	6.7	-9.5E-03	-3.6E-02	3.1E-02	0.56	6.03	
Architecture	CDIA-0	8	222.1	215.9-230.3	4.7	1.6E-02	3.4E-02	-1.8E-02	0.56	4.31	
Architecture	CDIA-0	10	168.5	167.5-176.7	8.7	4.9E-03	-4.6E-02	9.1E-04	0.56	7.8	
Architecture	CDIA-0	17	87.2	79.9-102.2	7	-7.2E-03	-3.7E-02	2.9E-02	0.56	6.31	
Architecture	CDIA-0	1A	66.4	48.2-71.9	6	1.4E-02	4.0E-02	-1.3E-02	0.56	5.42	
Architecture	FORM-0	4	72.7	72-83.5	10.2	1.6E+00	-2.4E+01	-6.9E+00	0.57	9.04	39.52
Architecture	FORM-0	5	96	91.1-109.6	7	4.1E+00	1.6E+01	-1.4E+01	0.57	6.28	
Architecture	FORM-0	8	222.1	215.9-230.3	5.1	-7.3E+00	-1.5E+01	8.4E+00	0.57	4.61	
Architecture	FORM-0	10	168.5	168-176.7	9	-2.3E+00	2.1E+01	-6.4E-01	0.57	8.07	
Architecture	FORM-0	17	87.2	79.9-102.2	7	3.2E+00	1.6E+01	-1.3E+01	0.57	6.28	
Architecture	FORM-0	1A	66.4	48.2-71.9	5.8	-6.0E+00	-1.7E+01	5.1E+00	0.57	5.24	
Architecture	CDIA-1	4	90.8	61.8-94.4	6.9	4.5E-02	3.6E-01	8.7E-02	0.53	6.21	31.87
Architecture	CDIA-1	5	95.7	94.1-97.8	14.6	-5.8E-02	-4.4E-01	3.5E-01	0.53	12.69	
Architecture	CDIA-1	10	169.5	165.3-197.7	6	1.6E-01	-2.9E-01	-1.5E-01	0.53	5.39	
Architecture	CDIA-1	17	15	5-18.4	8.5	4.7E-02	-4.3E-01	7.5E-02	0.53	7.57	
Architecture	FORM-1	4	90.8	61.8-94.4	6.9	-2.3E-03	-1.9E-02	-4.6E-03	0.53	6.24	32.03
Architecture	FORM-1	5	95.7	94.1-97.8	14.7	3.2E-03	2.4E-02	-1.9E-02	0.53	12.81	
Architecture	FORM-1	10	169.5	165.3-197.7	6	-8.5E-03	1.6E-02	8.3E-03	0.53	5.42	
Architecture	FORM-1	17	15	5-18.4	8.4	-2.4E-03	2.3E-02	-4.0E-03	0.53	7.56	
Architecture	CDIA-2	4	88.2	66.5-92.1	5.6	-4.6E-02	3.9E-01	1.9E-01	0.47	5.08	29.25
Architecture	CDIA-2	5	96.3	94.7-97.8	16.4	-1.7E-01	-5.5E-01	4.9E-01	0.47	14.16	
Architecture	CDIA-2	10	169.5	151.6-199.3	4.9	9.9E-02	-3.3E-01	-1.1E-01	0.47	4.42	
Architecture	CDIA-2	17	15.5	5-18.4	6.2	-3.9E-02	-4.3E-01	1.5E-01	0.47	5.6	
Architecture	FORM-2	4	88.2	66.5-92.1	5.6	1.6E-03	-1.3E-02	-6.3E-03	0.47	5.1	
Architecture	FORM-2	5	96.3	94.7-97.8	16.5	5.8E-03	1.9E-02	-1.6E-02	0.47	14.23	
Architecture	FORM-2	10	169.5	151.6-199.3	4.9	-3.4E-03	1.1E-02	3.8E-03	0.47	4.42	
Architecture	FORM-2	17	15.5	5-18.4	6.2	1.3E-03	1.4E-02	-5.2E-03	0.47	5.61	

Table 4.5
(continued)

Trait Class	Trait-Year	LG	Pos	CI	LOD	mu	a	d	h^2	Marker Vp (%)	Full Vp (%)
Biomass	HT-0	5	96.8	94.7-97.2	17.2	-8.1E+01	-2.1E+02	1.6E+02	0.42	14.81	19.35
Biomass	HT-0	6A	132.9	129-135.7	5	-4.1E+01	-1.2E+02	7.1E+01	0.42	4.53	
Biomass	DIA-1	5	96.3	94.1-97.8	27.8	-1.5E+00	-3.7E+00	3.1E+00	0.48	22.8	27.04
Biomass	DIA-1	8	176.6	161.5-178.6	4.7	-1.0E+00	-1.3E+00	1.7E+00	0.48	4.25	
Biomass	DVOL-1	5	96.3	94.1-96.9	30.1	-1.5E+00	-3.7E+00	2.9E+00	0.53	24.39	28.5
Biomass	DVOL-1	8	174.2	167.9-178.6	4.5	-1.1E+00	-1.2E+00	1.7E+00	0.53	4.11	
Biomass	MDDIA-1	2	39.2	34.7-44.7	4.3	-2.7E-02	-4.4E-02	3.1E-02	0.32	3.89	8.26
Biomass	MDDIA-1	5	95	92.5-100	4.8	-3.1E-02	-5.1E-02	3.1E-02	0.32	4.37	
Biomass	MDIA-1	2	38	27.4-48.9	5.1	-1.8E-02	-4.1E-02	2.4E-02	0.47	4.6	18.36
Biomass	MDIA-1	4	152.1	133.3-157.6	4.4	-2.3E-02	4.2E-02	2.8E-02	0.47	4.04	
Biomass	MDIA-1	5	95	92.5-97.2	11	-3.5E-02	-5.7E-02	5.2E-02	0.47	9.72	
Biomass	MDSA-1	2	39.2	34.7-44.7	4.3	-1.2E-02	-2.0E-02	1.4E-02	0.32	3.89	8.15
Biomass	MDSA-1	5	95	92-100	4.7	-1.3E-02	-2.2E-02	1.3E-02	0.32	4.26	
Biomass	MSA-1	2	38	27.4-49.9	4.8	-9.7E-03	-2.4E-02	1.2E-02	0.52	4.38	27.83
Biomass	MSA-1	4	152.1	133.3-157.6	4.6	-1.1E-02	2.6E-02	1.4E-02	0.52	4.19	
Biomass	MSA-1	5	96.5	92.5-97.4	12.1	-2.2E-02	-3.5E-02	3.3E-02	0.52	10.66	
Biomass	MSA-1	14	4.2	0-157.4	4.8	-8.6E-03	-2.3E-02	1.5E-02	0.52	4.38	
Biomass	MSA-1	6A	131.6	124.7-149.4	4.6	-7.2E-03	-2.6E-02	5.4E-03	0.52	4.22	
Biomass	MVOL-1	5	95	92.5-97.2	13.8	-2.5E-01	-4.8E-01	4.4E-01	0.51	12.07	17.71
Biomass	MVOL-1	6A	131.9	124.7-149.4	6.2	-6.1E-02	-3.9E-01	4.3E-02	0.51	5.64	
Biomass	MXSA-1	4	134.2	126.5-141.4	4.4	7.2E-03	3.4E-02	-1.9E-02	0.54	4	25.61
Biomass	MXSA-1	5	96.5	96.3-97.2	17.4	-3.4E-02	-6.5E-02	6.1E-02	0.54	14.96	
Biomass	MXSA-1	6A	132	124.7-138.3	7.4	-1.0E-03	-5.2E-02	-2.1E-03	0.54	6.65	
Biomass	MXVOL-1	5	96.5	94.7-97.2	17.3	-1.6E-01	-3.4E-01	3.0E-01	0.52	14.83	22.09
Biomass	MXVOL-1	6A	132	124.7-136.6	8.1	-3.5E-03	-2.8E-01	-3.9E-03	0.52	7.27	
Biomass	SA-1	5	96.3	94.1-97.8	29.3	-2.5E-01	-6.0E-01	5.0E-01	0.51	23.83	23.83
Biomass	SiDIA-1	5	94.7	92.5-104.7	9.8	-1.4E-02	-4.4E-02	2.4E-02	0.42	8.67	14.9
Biomass	SiDIA-1	6A	133.1	131-155.2	6.9	-7.8E-03	-4.0E-02	1.0E-02	0.42	6.23	

Table 4.5
(continued)

Trait Class	Trait-Year	LG	Pos	CI	LOD	mu	a	d	h^2	Marker Vp (%)	Full Vp (%)
Biomass	SiDW-1	5	94.7	92.5-104.7	10.8	-1.4E-02	-4.8E-02	2.3E-02	0.39	9.60	14.18
Biomass	SiDW-1	6A	139	125.9-155.2	5	-1.1E-02	-3.3E-02	1.6E-02	0.39	4.58	
Biomass	SiSSA-1	5	94.7	92.5-104.7	9.8	-3.8E-02	-1.1E-01	6.2E-02	0.42	8.67	14.9
Biomass	SiSSA-1	6A	133.1	131-155.2	6.9	-2.1E-02	-1.0E-01	2.6E-02	0.42	6.23	
Biomass	SiVOL-1	5	94.7	92.5-110.3	8.8	-2.2E-02	-6.4E-02	3.9E-02	0.4	7.85	13.18
Biomass	SiVOL-1	6A	133.1	130.7-155.2	5.9	-1.4E-02	-5.6E-02	2.2E-02	0.4	5.33	
Biomass	STNo-1	5	96.3	94.1-100	23.8	-2.7E-01	-6.4E-01	5.4E-01	0.44	19.87	24.22
Biomass	STNo-1	8	176.6	161.5-178.6	4.8	-1.6E-01	-2.9E-01	2.6E-01	0.44	4.36	
Biomass	VOL-1	5	96.3	94.1-97.2	28.1	-2.9E+00	-6.9E+00	5.6E+00	0.51	23.04	23.04
Biomass	HT-1.1	4	3.2	0-9.7	5.5	4.1E+03	6.8E+04	1.1E+04	0.57	4.99	35.97
Biomass	HT-1.1	5	99.6	87.7-108.5	10.9	-3.1E+04	-8.1E+04	7.3E+04	0.57	9.62	
Biomass	HT-1.1	8	205.9	201.8-227.8	8.6	-1.5E+04	7.8E+04	3.5E+04	0.57	7.67	
Biomass	HT-1.1	17	27.8	16.9-38.3	5.1	1.9E+04	-6.7E+04	-2.7E+04	0.57	4.65	
Biomass	HT-1.1	6A	132.4	129.8-135	10.2	-2.7E+04	-8.3E+04	6.1E+04	0.57	9.05	
Biomass	HT-1.2	3	237.8	235.9-239.3	4.8	-1.8E+02	-2.5E+02	3.8E+02	0.49	4.40	28.48
Biomass	HT-1.2	5	95	92.5-111	9.9	-2.0E+02	-4.5E+02	4.1E+02	0.49	8.80	
Biomass	HT-1.2	10	140.6	139.9-142.5	10.3	-9.4E+01	-4.8E+02	2.4E+02	0.49	9.16	
Biomass	HT-1.2	6A	132.1	124.7-149.4	6.8	-2.3E+01	-4.5E+02	5.6E+01	0.49	6.11	
Biomass	DIA-2	2	284.1	282.9-329.9	4.4	-8.1E-01	6.0E-01	1.2E+00	0.43	4.00	24.9
Biomass	DIA-2	5	94.7	94.1-97.8	25.2	-9.1E-01	-2.1E+00	1.7E+00	0.43	20.9	
Biomass	DVOL-2	2	284.1	282.9-329.9	4.4	-1.0E+00	7.0E-01	1.4E+00	0.47	3.97	26.9
Biomass	DVOL-2	5	94.7	92.5-96.9	28	-1.2E+00	-2.7E+00	2.1E+00	0.47	22.94	
Biomass	HT-2	5	96.3	92.5-109.6	10	-8.2E+02	-1.8E+03	1.4E+03	0.46	8.89	25.4
Biomass	HT-2	10	140.6	139.9-142.2	14.3	-6.6E+02	-2.2E+03	1.3E+03	0.46	12.46	
Biomass	HT-2	6A	22.2	17.2-149.4	4.4	-8.8E+02	-1.5E+03	1.3E+03	0.46	4.05	
Biomass	MDDIA-2	4	153	143.9-162.1	4.4	-4.1E-02	4.9E-02	3.6E-02	0.24	4.01	4.01
Biomass	MDIA-2	4	152.1	141.4-157.6	5.1	-1.0E-02	1.7E-02	8.2E-03	0.38	4.63	12.34
Biomass	MDIA-2	5	95	92.5-97.4	8.6	-1.6E-02	-1.7E-02	2.0E-02	0.38	7.71	

Table 4.5
(continued)

Trait Class	Trait-Year	LG	Pos	CI	LOD	mu	a	d	h^2	Marker Vp (%)	Full Vp (%)
Biomass	MDSA-2	4	153	143.9-162.1	4.4	-2.0E-02	2.5E-02	1.8E-02	0.23	3.97	3.97
Biomass	MSA-2	4	144	133.3-157.6	5.2	-1.6E-02	3.6E-02	6.3E-03	0.45	4.73	13.75
Biomass	MSA-2	5	95	92.5-97.4	10.2	-3.4E-02	-3.7E-02	4.2E-02	0.45	9.02	
Biomass	MVOL-2	4	143.9	133.3-157.6	5.1	-7.1E-02	1.9E-01	2.1E-02	0.48	4.66	23.95
Biomass	MVOL-2	5	95	92.5-97.2	11.8	-1.8E-01	-2.3E-01	2.4E-01	0.48	10.42	
Biomass	MVOL-2	10	154	149.3-162.2	4.8	-4.5E-02	-1.7E-01	-6.0E-02	0.48	4.35	
Biomass	MVOL-2	6A	132	124.7-155.2	5	-5.6E-02	-1.9E-01	-7.8E-03	0.48	4.52	
Biomass	MXDIA-2	5	95.7	94.1-97.4	16.9	-1.2E-02	-1.4E-02	1.5E-02	0.52	14.53	18.97
Biomass	MXDIA-2	6A	139	18.3-149.4	4.9	-7.6E-03	-9.1E-03	5.0E-03	0.52	4.44	
Biomass	MXSA-2	5	95.7	94.1-97.4	16.9	-5.1E-02	-6.0E-02	6.4E-02	0.52	14.53	
Biomass	MXSA-2	6A	139	18.3-149.4	4.9	-3.2E-02	-3.8E-02	2.1E-02	0.52	4.44	
Biomass	MXVOL-2	5	95.7	94.1-97.2	16.9	-8.4E-02	-1.1E-01	1.1E-01	0.54	14.57	23.82
Biomass	MXVOL-2	10	154	150.8-186.2	5.3	-2.6E-02	-7.2E-02	-2.0E-02	0.54	4.78	
Biomass	MXVOL-2	6A	146.4	17.2-155.2	4.9	-3.5E-02	-7.4E-02	1.5E-02	0.54	4.47	
Biomass	SA-2	2	284.1	283.3-329.9	4.7	-2.0E-01	1.4E-01	2.7E-01	0.45	4.28	26.04
Biomass	SA-2	5	94.7	94.1-97.8	26.4	-2.3E-01	-4.8E-01	4.0E-01	0.45	21.76	
Biomass	SiDIA-2	5	95.7	94.7-97.2	14.6	-1.7E-01	-2.4E-01	3.1E-01	0.46	12.69	17.46
Biomass	SiDIA-2	6A	132.4	124.7-155.2	5.3	-5.3E-02	-2.0E-01	5.3E-02	0.46	4.77	
Biomass	SiDW-2	5	95.7	94.1-97.8	17.2	-8.0E-02	-1.2E-01	1.4E-01	0.47	14.81	14.81
Biomass	SiSSA-2	3	33	0-35	4.2	1.8E-02	1.6E+00	-3.6E-01	0.46	3.80	
Biomass	SiSSA-2	5	95.7	94.7-97.2	14.6	-1.6E+00	-2.1E+00	2.7E+00	0.46	12.69	
Biomass	SiSSA-2	6A	132.4	124.7-155.2	5.3	-4.7E-01	-1.7E+00	4.7E-01	0.46	4.77	
Biomass	SiVOL-2	5	95.7	94.1-97.8	13.8	-1.5E-01	-1.9E-01	2.7E-01	0.47	12.03	16.49
Biomass	SiVOL-2	6A	132.4	124.7-149.4	4.9	-3.9E-02	-1.7E-01	3.2E-02	0.47	4.47	
Biomass	STNo-2	5	94.7	92.5-97.8	21.4	-2.6E-01	-7.1E-01	5.2E-01	0.42	18.03	21.98
Biomass	STNo-2	8	176.6	167.9-178.6	4.3	-1.5E-01	-3.2E-01	2.4E-01	0.42	3.94	
Biomass	VOL-2	2	284.1	283.3-329.9	4.7	-1.6E+00	1.0E+00	2.3E+00	0.47	4.26	25.36
Biomass	VOL-2	5	94.7	94.1-97.8	25.5	-1.8E+00	-3.8E+00	3.1E+00	0.47	21.11	

Table 4.5
(continued)

Trait Class	Trait-Year	LG	Pos	CI	LOD	mu	a	d	h^2	Marker Vp (%)	Full Vp (%)
Composition	DEN-1	2	95	93.7-102.5	6	-1.2E-03	5.0E-03	1.5E-03	0.37	5.40	11.42
Composition	DEN-1	1A	76	69.6-103	6.7	-1.8E-05	5.3E-03	-1.3E-03	0.37	6.02	
Composition	CLS-2	2	369.9	366.2-369.9	4.9	-2.2E+00	-1.4E+01	-9.7E-01	0.53	4.48	25.17
Composition	CLS-2	8	174.6	171.2-178.6	7.9	-1.9E+00	-1.8E+01	-1.7E+00	0.53	7.12	
Composition	CLS-2	10	129	123.3-134.4	4.5	-5.8E+00	-1.3E+01	4.9E+00	0.53	4.13	
Composition	CLS-2	1A	59.3	51.7-75.6	5.6	7.1E-02	1.3E+01	-7.5E+00	0.53	5.12	
Composition	CLS-2	6A	11.6	0-29.7	4.8	-9.2E+00	-1.1E+01	1.2E+01	0.53	4.33	
Composition	DEN-2	2	131	50.9-133.2	6.2	-6.0E-03	1.9E-02	1.2E-02	0.61	5.60	35.3
Composition	DEN-2	3	42.8	36.7-50.7	4.5	-5.6E-03	-1.7E-02	1.1E-02	0.61	4.12	
Composition	DEN-2	5	94	91.1-104.7	10.6	-7.1E-03	-2.6E-02	1.4E-02	0.61	9.38	
Composition	DEN-2	14	54.3	35.1-71.6	4.5	1.2E-03	1.9E-02	-1.7E-03	0.61	4.10	
Composition	DEN-2	16	136	132.7-139.5	5.5	1.6E-03	1.9E-02	-6.9E-03	0.61	4.95	
Composition	DEN-2	1A	72.1	69.2-78.1	8	2.4E-03	2.3E-02	-6.7E-03	0.61	7.15	
Composition	HCL-2	2	48.1	16.3-136.4	4.7	-6.0E-07	-4.1E-06	1.2E-08	0.51	4.25	19.5
Composition	HCL-2	3	52	8.5-56.8	5.8	-2.8E-07	4.8E-06	-7.1E-07	0.51	5.27	
Composition	HCL-2	15	95.3	92.4-102	5.8	6.8E-07	-3.8E-06	-1.9E-06	0.51	5.22	
Composition	HCL-2	18	115.9	105.6-116.7	5.2	-2.0E-06	3.5E-06	3.1E-06	0.51	4.76	
Composition	LIG-2	8	164.2	154.1-176.9	5.9	1.2E-05	3.5E-05	-4.4E-06	0.38	5.35	5.35
Foliar	LFA-1	5	96.7	95.7-97.4	29.1	-4.8E-02	-1.2E-01	7.8E-02	0.63	23.72	43.66
Foliar	LFA-1	10	156.2	140.2-158.6	5.2	-1.4E-04	-5.0E-02	-2.3E-02	0.63	4.76	
Foliar	LFA-1	14	92.2	61.7-112.3	6.7	-1.2E-02	-6.4E-02	7.2E-03	0.63	6.02	
Foliar	LFA-1	15	113	79.4-115.8	4.9	-2.1E-02	-4.9E-02	1.5E-02	0.63	4.45	
Foliar	LFA-1	6A	152	124.7-155.9	5.2	2.7E-03	-5.4E-02	-1.8E-02	0.63	4.71	
Foliar	LFDW-1	5	96.5	94.7-97.4	21.4	-1.8E-02	-3.3E-02	2.7E-02	0.52	18.05	30.28
Foliar	LFDW-1	14	93	61.3-111.3	6	-4.2E-03	-2.1E-02	1.9E-04	0.52	5.42	
Foliar	LFDW-1	15	104.7	100-115.8	7.6	-7.5E-03	-2.0E-02	4.7E-03	0.52	6.82	
Foliar	LFF-1	2	244	199.4-245.1	8.7	4.9E-04	-3.1E-03	-7.5E-04	0.49	7.77	20.14
Foliar	LFF-1	5	96.9	90.1-99.6	4.7	2.0E-04	-2.2E-03	-5.1E-05	0.49	4.32	

Table 4.5
(continued)

Trait Class	Trait-Year	LG	Pos	CI	LOD	mu	a	d	h^2	Marker Vp (%)	Full Vp (%)
Foliar	LFF-1	14	62.3	57.9-125.3	4.4	1.1E-04	-2.2E-03	3.1E-04	0.49	4.01	
Foliar	LFF-1	15	122.4	115.8-125.6	4.4	3.2E-04	2.1E-03	-7.8E-04	0.49	4.04	
Foliar	LFL-1	2	354.8	197.7-369.9	5.2	-2.7E-04	8.7E-03	-4.1E-03	0.54	4.73	38.6
Foliar	LFL-1	5	96.5	94.1-97.2	12.4	-7.1E-03	-1.3E-02	1.0E-02	0.54	10.93	
Foliar	LFL-1	15	114	107.8-119.1	7	-4.5E-03	-9.7E-03	5.1E-03	0.54	6.29	
Foliar	LFL-1	18	53.4	44.3-57.7	4.7	1.5E-03	6.2E-03	-8.2E-03	0.54	4.24	
Foliar	LFL-1	13A	21	7.1-32	6.7	-2.4E-03	-1.1E-02	7.1E-04	0.54	6.00	
Foliar	LFL-1	6A	126.4	121.4-127.7	7.1	2.9E-03	-8.9E-03	-1.0E-02	0.54	6.41	
Foliar	LFP-1	2	354.8	200.3-356.6	4.9	-2.9E-05	2.6E-04	-5.4E-05	0.47	4.45	29.06
Foliar	LFP-1	5	96.5	94.1-97.4	10.2	-2.0E-04	-3.2E-04	2.9E-04	0.47	9.02	
Foliar	LFP-1	15	111	104.6-121.7	6.8	-8.4E-05	-2.9E-04	7.5E-05	0.47	6.09	
Foliar	LFP-1	18	53.4	40.9-61.7	4.3	4.8E-05	1.8E-04	-2.3E-04	0.47	3.96	
Foliar	LFP-1	13A	15.5	10.2-26.6	6.1	-7.7E-05	-2.9E-04	5.5E-05	0.47	5.54	
Foliar	LFR-1	2	198	195.5-240.6	8.9	7.2E-03	1.1E-01	-1.7E-02	0.62	7.97	34.48
Foliar	LFR-1	5	97.4	95.7-98.8	5.9	3.7E-03	8.6E-02	-1.8E-02	0.62	5.33	
Foliar	LFR-1	10	114.1	103.4-126.5	4.6	-1.6E-02	-8.2E-02	2.8E-02	0.62	4.15	
Foliar	LFR-1	14	60	56.9-74.1	8.1	-2.1E-02	1.0E-01	2.4E-02	0.62	7.29	
Foliar	LFR-1	13A	8.4	5.9-32	6	-1.3E-02	-8.9E-02	5.3E-03	0.62	5.42	
Foliar	LFR-1	13C	12	0-14.9	4.7	9.7E-03	7.7E-02	-4.5E-02	0.62	4.32	
Foliar	LFW-1	5	96.8	95.7-97.8	29.6	-7.4E-03	-2.1E-02	1.2E-02	0.68	24.06	37.99
Foliar	LFW-1	10	156.2	152.5-206.6	5	-4.7E-04	-8.6E-03	-2.6E-03	0.68	4.58	
Foliar	LFW-1	14	62.3	58.7-74.1	10.6	-2.1E-03	-1.4E-02	2.2E-03	0.68	9.35	
Foliar	LFA-2	4	9.7	0.6-11	4.8	-1.6E-02	-1.1E-02	4.5E-02	0.47	4.40	25.7
Foliar	LFA-2	5	97	94.7-100.7	11.8	-1.1E-02	-4.9E-02	2.8E-02	0.47	10.38	
Foliar	LFA-2	9	89.4	88.6-102.2	6.4	4.3E-03	-3.9E-02	-2.6E-03	0.47	5.78	
Foliar	LFA-2	7A	69	60.6-76	5.7	1.7E-02	3.2E-02	-2.6E-02	0.47	5.15	
Foliar	LFDW-2	2	150.4	147.4-167.1	6.1	3.9E-03	2.4E-02	-1.0E-02	0.49	5.51	21.57
Foliar	LFDW-2	5	96.9	94.7-100.7	13.8	-1.3E-02	-3.3E-02	2.2E-02	0.49	12.05	

Table 4.5
(continued)

Trait Class	Trait-Year	LG	Pos	CI	LOD	mu	a	d	h^2	Marker Vp (%)	Full Vp (%)
Foliar	LFDW-2	7A	69.1	65.4-70.4	4.4	5.0E-03	1.9E-02	-1.2E-02	0.49	4.02	
Foliar	LFF-2	2	301	274.7-339.2	7.3	4.5E-03	-5.1E-03	-4.5E-03	0.49	6.57	10.67
Foliar	LFF-2	10	93.7	61-98	4.5	2.6E-03	5.0E-03	-1.5E-04	0.49	4.10	
Foliar	LFL-2	5	108	90.1-111.9	6.8	-6.6E-03	-1.2E-02	1.1E-02	0.36	6.09	12.16
Foliar	LFL-2	9	95.1	88.9-103.3	6.7	-3.6E-03	-1.4E-02	5.4E-03	0.36	6.06	
Foliar	LFP-2	5	100	90.6-110	7.1	-3.8E-03	-6.2E-03	6.5E-03	0.35	6.38	12.59
Foliar	LFP-2	9	97.1	88.9-100.2	6.9	-1.9E-03	-7.4E-03	2.5E-03	0.35	6.21	
Foliar	LFR-2	2	80	34.7-100.7	6.8	-1.4E-02	3.1E-02	1.6E-02	0.59	6.16	26.14
Foliar	LFR-2	3	75.9	69.5-81.3	5.3	-1.3E-02	-2.5E-02	1.4E-02	0.59	4.79	
Foliar	LFR-2	10	260	246.3-262.2	8.2	-7.1E-03	3.5E-02	1.2E-03	0.59	7.33	
Foliar	LFR-2	14	58.4	34.8-88.7	4.3	-9.6E-03	2.5E-02	5.3E-03	0.59	3.96	
Foliar	LFR-2	1A	8	3.8-94.6	4.3	-1.6E-02	-2.2E-02	2.0E-02	0.59	3.90	
Foliar	LFW-2	5	97.2	95.3-97.8	16.1	-3.5E-03	-2.6E-02	1.1E-02	0.65	13.90	29.69
Foliar	LFW-2	9	89.4	84.8-98.6	4.7	4.8E-03	-1.4E-02	-4.3E-03	0.65	4.25	
Foliar	LFW-2	10	212.1	205.4-262.2	7.1	2.3E-03	-1.8E-02	7.4E-04	0.65	6.40	
Foliar	LFW-2	7A	69	65.5-80.1	5.7	9.2E-03	1.4E-02	-1.1E-02	0.65	5.15	
Pathology	RUST-1	5	94	92.5-95.3	15.4	-6.3E-03	-8.6E-03	1.6E-02	0.62	13.34	37.43
Pathology	RUST-1	10	150.8	149.3-155.7	14.4	1.6E-03	1.3E-02	-1.7E-03	0.62	12.54	
Pathology	RUST-1	1A	44.8	36.8-45.4	13.2	-2.3E-03	-1.2E-02	6.3E-03	0.62	11.55	
Pathology	RUST-2	1A	40	39.3-72.7	4.7	-9.1E-05	-6.2E-04	2.5E-04	0.24	4.29	4.29
Physiology	SEX	15	68.7	64.3-65.2	70	1.6E-16	-9.3E-17	-4.1E-17	0.68	45.01	45.01
Physiology	STC-0	17	84.8	79.9-85.2	78.9	-1.8E-01	-9.7E-01	4.5E-01	0.69	51.99	51.99
Physiology	SLA-1	8	207.7	185.7-213.4	5.3	1.1E-04	-6.4E-04	8.6E-05	0.36	4.80	4.8
Physiology	SLA-2	2	160.3	158.1-172.9	6.3	1.4E-02	-2.0E-02	-1.2E-02	0.3	5.68	9.75
Physiology	SLA-2	4	132	108.6-143.3	4.5	9.4E-04	-1.5E-02	1.3E-02	0.3	4.07	
Physiology	SPAD-2	2	175.5	173.8-230.9	5.6	-1.1E+02	1.9E+02	8.4E+01	0.42	5.04	18.93
Physiology	SPAD-2	10	142.5	137.5-146.5	11	1.6E+00	2.4E+02	-1.5E+02	0.42	9.76	

LOD support intervals for nearly all biomass-related trait QTL overlapped considerably on chr05 (Fig. 2), where the cumulative physical support interval spaned ~2 Mb and contains a mere 30-40 gene models. Three adjacent markers on chr05 (~0.5 cM genetic, ~1 Mb physical) alternate peak LOD values for DIA, SA, VOL, and STNo, each individually contributing 15-25% Vp. Biomass trait averages (opposed to total), MDIA-1, MSA-1, and MVOL-1, as well as for HT-1.1 and HT-1.2, all had overlapping QTL LOD support on chr06. In addition, QTL for measurements made on single representative stem segments for composition analysis, SiDIA and SiSSA, co-located with biomass averages for both years on chr06.

On average, foliar traits had relatively high heritability in the F_2 family, which ranged from 0.45 to 0.64. However, in 2016, a severe drought considerably impacted foliar traits, shrinking average LFA by 10%. Seasonal differences of F_2 foliar traits were further pronounced by the relative lack of reproducible, overlapping LOD intervals for the same trait between years. However, QTL clustered to linkage groups quite well within year, especially for LFDW, LFA, LFL, and LFP. In addition, QTL for foliar traits were more dispersed throughout the genome than other trait classes, with LOD support on a total of nine chromosomes.

For wood composition QTL, overlapping LOD intervals were identified on chr01 for CLS-2, LIG-2, DEN-1, and DEN-2, chr02 for DEN-1 and DEN-2, chr03 for HCL-2 and DEN-2, and chr08 and chr10 for CLS-2 and LIG-2. Unique QTL for HCL-2 were identified on chr15, as well as chr18, which was the only trait that had LOD support on the chromosome in the F_2.

On average, establishment-year FORM was greater (50° ± 0.4°) than both first (45° ± 0.2°) and second (46° ± 0.2°) year post-coppice measurements. Pearson correlations (r^2) for CDIA-0 and CDIA-1 (r^2 = 0.49) and CDIA-2 (r^2 = 0.48) were improved using the *arctan2* (FORM) transformation, with r^2 values of 0.60 and 0.59, respectively. High first and second year

post-coppice correlations of CDIA (r^2 = 0.84) and FORM (r^2 = 0.87) reflect the repeatability of QTL for these traits on chromosomes 4, 5, 10, and 17, and unique QTL for CDIA-0 and FORM-0 identified on chr08.

Stem architecture traits were most correlated with STC-0, especially for CDIA-0 (r^2 = 0.25, $P < 1\times10^{-6}$), such that plants with wider branching architecture were more likely to have either intermediate or red STC-0, compared to those with narrow stools. On average, CDIA-0 was 20 cm, 26 cm, and 30 cm, for green, intermediate, and red STC-0 classes, respectively. The greatest marker LOD obtained for any trait was for STC-0 (LOD = 76.9; Vp = 51%). The support interval for the STC-0 QTL represented a physical interval ~225 Kb on chr17, encompassing less than 20 unique gene models. The grandparents of the F_2, 94006 and 94001, exhibited intermediate and red STC-0, respectively, whereas the parents of the F_2, 'Wolcott' and 'Fish Creek', both had intermediate STC-0 classes. STC-0 among the F_1 siblings of 'Wolcott' and 'Fish Creek', gave a ratio 1:3 for intermediate:red, and only intermediate and red STC-0 classes appeared in the F_1. Conversely, the F_2 family ratio was 1:4:6 for green:intermediate:red STC-0 classes, respectively.

Accounting for a total of 22% Vp, two minor and one major QTL for SPAD-2 were identified on chromosomes 2, 4, and 10, respectively.

Willow leaf rust severity (%) was scored in late-season 2015 (RUST-1) and 2017 (RUST-2), with the former exhibiting greater genetic variance, which provided more power to detect QTL, compared to the latter. For RUST-1, associated QTL support intervals were identified on chromosomes 1, 5, and 10. While RUST-2 was underpowered, the support interval for a QTL on chr01 co-located with that of RUST-1, albeit with a comparatively wider interval.

4.4.3 Concordance between GWAS and Linkage Mapping

Overall, there were only a handful of significant associations ($p < 1\times10^{-6}$) from GWAS identified within LOD support intervals of QTL corresponding to the same traits in the F_2 family (Table 4.6). Markers associated with VOL for the PTL trial were within the LOD support interval for HT-1 on chr17 in the F_2, whereas associations with VOL in the Geneva and WVU trials were located within the chr05 interval for HT-1 in the F_2. Internode length (INLEN) and VPHE were also within a support interval for HT-1, but for a QTL located on chr04. Associations for LFW in the Portland trial were within chr02 support intervals of QTL for the F_2 foliar traits: LFR, LFF, and SLA. Specific leaf area (SLA) and SPAD1 measured in the Portland association trial overlapped with chr10 support intervals of QTL for LFR and LFF in the F_2. In addition, GWAS hits for SPAD2 were found within intervals for LFF on chr02 and STC on chr17. Significant GWAS hits for the composition traits, CLS and LIG, at the WVU location were positioned within the chr10 support intervals of QTL for VOL and DVOL (wood density × stem volume) in the F_2, whereas those for ASH at the Geneva location were within the chr08 support intervals of QTL for LIG and CLS. Floral phenology (FPHE) scored at both Geneva and WVU locations and LFF in the Portland location each had significant hits within the chr15 support interval for SEX in the F_2.

Table 4.6 Subset of significant associations (p-value < 1×10^{-6}) in the *S. purpurea* association panel positioned within the F_2 LOD support intervals (LOD > 4) of QTL for related traits.

GWAS	Location	QTL	Chr	GWAS Position	QTL Physical Interval[1]
Biomass					
VOL-1	PTL	HT-1.1	17	7295573	2814640-8046575
VOL-2	GVA	HT-1.1	5	7326392	7158399-12415317
VOL-2	WVU	HT-1.1	5	7841126	7158399-12415317
VOL-2	PTL	HT-1.1	17	7295573	2814640-8046575
INLEN-1	GVA	HT-1.1	4	1514308	1262802-14842501
INLEN-1	GVA	HT-1.1	4	7793610	1262802-14842501
Foliar					
LFW-1	PTL	LFR-1	2	3093644	2616294-7022046
LFW-1	PTL	LFF-2	2	3093644	1589349-12550749
LFW-1	PTL	LFF-1	2	3093644	2260613-7022046
LFW-1	PTL	SLA-1	2	3093644	2260613-7022046
LFF-1	PTL	SEX	15	3303898	2376419-6573289
SLA-2	PTL	LFR-1	10	3305469	1075618-16060177
SLA-2	PTL	LFR-1	10	3305466	1075618-16060177
SLA-2	PTL	LFR-1	10	3305474	1075618-16060177
SLA-2	PTL	LFF-2	10	3305466	828545-9870491
SLA-2	PTL	LFF-2	10	3305469	828545-9870491
SLA-2	PTL	LFF-2	10	3305474	828545-9870491
Composition					
ASH-2	GVA	ASH-2	8	7437295	5420927-7720138
ASH-2	GVA	CLS-2	8	7437295	5420927-7720138
CLS-1	WVU	DVOL-1	10	12108038	10231740-15628389
CLS-1	WVU	DVOL-2	10	12108038	10231740-15628389
LIG-1	WVU	DVOL-1	10	12108038	10231740-15628389
LIG-1	WVU	DVOL-2	10	12108038	10231740-15628389
Physiology					
SPAD-1-1	PTL	LFW-2	10	15971782	13820753-16217863
SPAD-1-1	PTL	LFR-1	10	15971782	1075618-16060177
SPAD-1-1	PTL	LFR-2	10	15971782	15646582-16120119
SPAD-1-1	PTL	STC-0	17	10293787	9257004-10667656
SPAD-2-2	PTL	LFF-2	2	10364274	1589349-12550749
SPAD-2-2	PTL	LFF-2	2	10364291	1589349-12550749
Phenology					
FPHE-2	GVA	SEX	15	4068064	2376419-6573289
FPHE-2	WVU	SEX	15	5609619	2376419-6573289
VPHE-2	WVU	HT-1.1	4	4632079	1262802-14842501

[1]Estimated QTL physical interval derived from marker mappings flanking LOD support intervals.

4.5 Discussion

4.5.1 Linkage Disequilibrium

For association mapping, a large number of GBS markers were generated using a single, methylation-sensitive restriction enzyme; however, the density of markers was relatively low given the size of the *S. purpurea* reference genome (~350 Mb). Marker density may affect mapping resolution by limiting the power for complete dissection of complex trait architectures if small effect polymorphisms are located far from regulatory gene space, but may be improved upon with the use of double-digestion with two restriction enzymes. In addition, the sample size of the association panel was the most likely factor in limited associations identified. Nevertheless, a number of significant genome-wide SNPs were detected in GWAS, despite the small sample size, and rather stringent Bonferroni threshold used.

Multiple significant associations were found which could be due to longer stretches of LD on chromosomes not previously detected, as LD between distant loci can inflate single locus test statistics (Thomas et al., 2011). Many highly heterozygous, outcrossing plants, including some tree species, display rapid LD decay which was reported around 2.6 kb for *P. deltoides* and greater than 1 kb in *P. trichocarpa* (Slavov et al., 2012), compared to self-pollinated species, such as rice (75-150 kb) (Huang et al., 2010) or *Arabidopsis thaliana* (250 kb) (Nordborg et al., 2002). Despite this complexity, this suggests that alternative SNPs may be located in the proximity to a significantly associated SNP even though that marker itself might not be causative. It should be noted that some of the genetic architectures of the traits in this study are highly complex, where tens or hundreds of causative polymorphisms with minor effects may exist genome-wide.

4.5.2 QTL Hotspots in the F_2

The majority of biomass-related QTL identified on chromosomes 4, 5, 6, and 10 showed significant overlap of LOD support intervals, especially for those identified on chr05 and chr10. A number of genes within the chr05 QTL hotspot have functions related to vegetative phase change and meristem maintenance, such as a homeobox-leucine zipper protein (HB-16), squamosa promoter-binding like transcription factor (SPL16), MADS box interactor-like protein (MIP1), helix loop helix DNA-binding transcription factor (bHLH74), GATA transcription factor (ZIM), GLABRA 2 expression modulator (GEM1), and two AP2/ERF transcription factors (AP2 and ANT). Mutants in these genes have led to altered flowering time response to photoperiod, cell expansion and elongation, as well as defects in apical dominance and branching architecture (Shikata et al., 2004; Preston and Hileman, 2013). One of the primary candidates within the chr05 hotspot is a squalene monooxygenase/epoxidase gene. Squalene epoxidase (SQE1) catalyzes the first oxygenation step in sterol biosynthesis, which is a part of terpene metabolism. Triterpenoid biosynthetic mutants often display severe pleiotropic defects (Rasbery et al., 2007). For instance, Arabidopsis SQE1 mutants incited hypersensitivity to drought stress, altered stomatal response, and defects in root architecture, due to mislocalization of a ROS-generating, respiratory burst oxidase homolog, leading to altered ROS production (Pose et al., 2009, Marino et al., 2012). It is feasible that altered SQE1 activity may explain the pleiotropic nature of the chr05 QTL hotspot and MAS at this locus could serve as an early biomass indicator.

In poplar, five main biomass QTL hotspots were identified on linkage groups representing physical chromosomes 3, 4, 10, 14, and 19 (Rae et al., 2009). Particularly interesting was the QTL hotspot on the end of *Populus* chr10 (PBL-3), which had LOD support

for many biomass-related traits common to this study, especially QTL for HT, VOL, and LFA, which also mapped to *Salix* chr10. In addition, using a F_1 *S. viminalis* × *S. schwerinii* family, Ghelardini et al. (2014) identified QTL for vegetative phenology traits on chr10, but had limited power to detect large-effect QTL or make gene-level inferences, due to low marker coverage. For the reason many biomass-related traits in the F_2 were highly correlated or autocorrelated (e.g., HT, DIA, SA, VOL, STNo, etc.), it is expected that these traits share common genetic architectures. Yet, the chr05 hotspot contains QTL for numerous allometric and non-allometric traits, so it is likely that these QTL represent either high-effect *cis*-acting QTL of a major regulator or tight linkage of many important genes. Both the former and latter seem plausible, because this region contains multiple genes known to be key players in developmental and metabolic pathways. *Trans*-effects (e.g., pleiotropy) are more likely to be subject to purifying selection, and may be why these hotspots were not observed in GWAS. While selection for beneficial alleles at QTL hotspots will likely result in improved yield, realistically, the greatest gain from selection would be realized for related individuals.

4.5.3 Foliar

Although sex was included as a covariate, the QTL LOD support intervals for LFDW-1, LFA-1, LFL-1, and LFP-1 were located within a pseudoautosomal region (PAR), which flanks the large, non-recombining SDR on *Salix* chr15.

4.5.4 Wood Chemical Composition

The peak LOD marker for CLS-2 on chr01 is within the coding sequence of a UDP-glucose 4-epimerase (UGE5), which functions as a catalyst for the interconversion between UDP-glucose and UDP-galactose in galactose metabolism (Rosti et al., 2007). Neighboring genes residing within LOD support intervals for CLS-2 on chr08 encode for endo-1,4-β-

glucanase and phosphoglucomutase, both of which are involved in carbohydrate biosynthesis. Likewise, the peak LOD marker for LIG-2 is located within the coding sequence of a chitinase-like class I glycoside hydrolase 19 family protein (CTL2). Expressed primarily in xylem and interfasicular fibers, CTL2 has been shown to be required for cell wall biosynthesis and lignin accumulation in Arabidopsis mutants (Hossain et al., 2010). Importantly, a laccase gene, similar to Arabidopsis IRREGULAR XYLEM 12 (IRX12), functions in lignin degradation and secondary xylem cell wall lignification (Brown et al., 2005).

4.5.5 Architecture

In the association panel, best hits for CDIA were located in the coding sequence for a glycoside hydrolase gene, GH9, which is the second largest cellulase gene family (Davies and Henrissat, 1995). Functional work with this enzyme family indicate that they are involved in cell wall modification during fruit softening, abscission, growth, and wood formation (Urbanowicz et al., 2007; Du et al., 2015). Protein homologs for this gene matched other known GH9 proteins in *Populus trichocarpa* and *Theobroma cacao*. Based on sequence alignments and Pfam database searches, the candidate gene found in *S. purpurea* belongs to subclass B, which comprises secreted proteins with only one catalytic domain (Urbanowicz et al., 2007). GH9B, synonymous with endo-1,4-β-glucanase 11, has been reported with activities for cello-oligosaccharide release and xyloglucan cleavage in plants, but the GH9 superfamily has yet to be characterized in *Salix*. One well-studied GH9 clade includes a membrane-associated endoglucanase (KORRIGAN) that is part of the cellulose synthesis complex and that influences the organization of cellulose in the wall (Bhandari et al., 2006). Shrub willows are comprised of numerous woody stems that exhibit various sweeping branching angles from the base of the crown, which requires expanding cell walls to withstand high tensile forces generated by cell wall stress. This gene may have future

implications on targeting mechanical properties of wood to influence biomechanical strength and therefore directly influencing plant architecture.

For all CDIA and FORM measurements, a common QTL on chr04 centers on a gene encoding for a lateral organ boundaries (LOB) domain (LBD) protein (LBD40), synonymous with an ASYMETRIC LEAVES 2-like protein. Expressed at the base of initiating lateral organs, LBD proteins have been implicated in the regulation of anthocyanin and nitrogen metabolism (Majer and Hochholdinger, 2011), auxin signaling (Fan et al., 2012), and defense responses to pathogens (Thatcher et al., 2012). What was unique to the F_2, architectural traits were also measured on establishment-year growth. QTL for CDIA-0 and FORM-0, had LOD support on chr17, which overlapped with the support interval for STC-0. A kanadi-like transcriptional repressor (KAN1) lies within the chr17 interval. KAN1 has been shown to regulate both lateral organ identity by repressing ASYMETRIC LEAVES 2 (AS2) (Wu et al., 2008a), and axial cell elongation and xylem differentiation in the cambium by repressing the auxin efflux carrier, PIN-FORMED1 (PIN1) (Ilegems et al., 2010). For the reason QTL for establishment-year stem architecture measurements co-localized with fewer QTL for later years, it is likely this phenotype is transcriptionally regulated.

4.5.6 Physiology

Candidate genes for STC-0 positioned within the chr17 support interval are those encoding for GDP-L-galactose phosphorylase (GGP), L-Galactono-1,4-lactone dehydrogenase (GLDH), and glutathione S-transferase/dehydroascorbate reductase 3 (DHAR3), as well as six Myb transcription factors. Both GGP and GLDH are major players in the Smirnoff-Wheeler pathway (Smirnoff and Wheeler, 2000), which is the primary route to L-ascorbate (vitamin C) biosynthesis in plants. GDP-L-galactose phosphorylase is an enzyme that catalyzes the first

committed step in L-ascorbate biosynthesis, whereas GLDH catalyzes the terminal step (Linster and Clarke, 2008). Dehydroascorbate reductase (DHARs) genes are key components in the ascorbate recycling system and are involved in scavenging of ROS under oxidative stress (redox homeostasis) (Dixon et al., 2002), and also serve as a substrate for oxidized anthocyanins (Dixon et al., 2011). Of the six Myb transcription factors within the LOD support interval for STC-0, four are paralogous to Arabidopsis PRODUCTION OF ANTHOCYANIN PIGMENT 1 (PAP1) and two are paralogous to PRODUCTION OF ANTHOCYANIN PIGMENT 2 (PAP2). Transgenic *P. trichocarpa* plants overexpressing MYB119/PAP1 showed elevated accumulation of anthocyanins in leaf, stem, and root tissues, without significantly affecting normal growth (Cho et al., 2016). The sheer physical genomic proximity of GGP, GLDH, and DHAR3 on *Salix* chr17 suggests that the expression of these genes may be regulated by a common *trans*-factor, and that *cis* differences among haplotypes could explain variation in L-ascorbate activity and subsequent effect on STC-0.

For the reason STC-0 was collected mid-summer, a meager establishment-year canopy (or lack thereof) likely permitted high-light or UV-B radiation greater access to developing stems. Further, the accumulation of pigments was far more pronounced along south stem faces, compared to those masked by foliage, so anthocyanin accumulation may be limited by ascorbate deficiency. If the accumulation of anthocyanin is proportional to mRNA accumulation in response to oxidative stress, then the role of GGP and GLDH in L-ascorbate biosynthesis in *S. purpurea* may serve to regulate photoprotectant activity by quenching ROS under high-light conditions. If desired, reliable selection for STC-0 could be easily accomplished via MAS. Positioned within the chr02 LOD support interval in the F_2, the best GWAS hit for SPAD-2 was nearest to a Myc-like anthocyanin regulatory protein, GLABRA 3 (GL3). Induced by nitrogen

deficiency and UV light, GL3 has been implicated in epidermal cell-fate specification (Payne et al., 2000), and associating with PAP1 and PAP2 in anthocyanin biosynthesis (Zhang et al., 2003). This gene may provide a link between oxidative stress and the regulation of STC and SPAD traits in *S. purpurea*.

One likely candidate for SPAD-2 within the chr10 QTL support interval in the F_2 is a cytoplasmic CTP synthase/UTP-ammonia ligase. Acting as a catalyst in the rate-limiting step of *de novo* CTP biosynthesis, which is the last committed step in pyrimidine metabolism, CTP synthase interconverts UTP to CTP using L-glutamine or ammonia as the source of nitrogen. Recently, it has been shown that CTP synthase forms filamentous structures, termed cytoophidia, which are predicted to function in enzymatic storage and compartmentalization. Highly sensitive to the cellular metabolic state (Aughey and Liu, 2016), the reduction of CTP synthase filament formation is thought to be induced by nutrient restriction. Another candidate for SPAD-2 is an isocitrate dehydrogenase (clCDH) gene located within the chr10 LOD support interval. Thought to supply 2-oxoglutarate for amino acid biosynthesis and ammonia assimilation, clCDH produces NADPH to promote redox signaling in response to oxidative stress. For instance, loss of clCDH function in Arabidopsis promoted H_2O_2-induced lesions in leaves, which was likely due to a suboptimal reductant supply of NADPH oxidases (Mhamdi et al., 2010). Defects in clCDH could impair redox signaling in *S. purpurea* leaves, leading to early senescence, and lower SPAD values.

4.5.7 Phenology and Sex

While it is common that male willows flower before females (Cronk et al., 2015), this study provides evidence that after the onset of flowering, male *S. purpurea* transition to vegetative growth (precocity) at a faster rate than females. One explanation for the difference in

the rate of precocity in *S. purpurea* is that developing seed in female catkins require more resources prior to seed maturation and dispersal, whereas males reallocate resources shortly after anther dehiscence and abscission. Nevertheless, it is unclear whether the rate of precocity or phenological sex dimorphism in *S. purpurea* significantly impacts the accumulation of biomass over time.

A strong association with FPHE was observed with a significant association peak on chr15 that fell within genes related to multiple protein binding domains. Significant GWAS hits for FPHE were within coding regions of a serine/threonine protein kinase 3 (STK3) and an ATP-dependent ClpB heat shock protein (HSP101). Both genes are located within the non-recombining SDR on *Salix* chr15. While serine/threonine protein kinases fulfill diverse roles in signaling and response to biotic and abiotic stressors, heat shock proteins (HSPs) are essential proteins that function to maintain cellular proteostasis by limiting the accumulation of protein aggregates induced by heat shock and confer thermotolerance. For instance, transgenic tobacco and cotton lines overexpressing AtHSP101 exhibited higher germination rates and greater pollen tube elongation in elevated temperatures and after exposure to heat shock (Burke and Chen, 2015). In AtHSP101 knockouts, mutant lines displayed defects in translational recovery and proper disassociation of protein aggregates after exposure to heat shock (Merret et al., 2017). Importantly, the same *Salix* HSP101 (SapurV1A.1386s0030) was differentially expressed ($-\log_{10}$ (*p*-value) = 10.3, $\log_2$ (female / male) = 3.1) between the shoot-tip transcriptome of full-sib F_1 *S. purpurea* female and male individuals (Carlson et al., 2017; see Chapter 5).

In female *S. purpurea*, catkins are maintained until after seed ripening and dispersal, whereas male aments fall off soon after flowering, so early protection from heat-induced programmed cell death could be critical for female reproductive success. Dioecy could possibly

have evolved to limit the accumulation of detrimental variation in regulators of floral phenology, which may have served as an early sexual antagonist in the evolution of dioecy in *S. purpurea*. However, throughout the evolution of dioecy in *Salix*, preferential selection for male precocity may simply have been a consequence of sexual dimorphism, since selection for early ovules is intense.

Previous publications have pointed to a large, non-recombining region on chr15 as the SDR for *Salix* (Hou et al., 2015; Pucholt et al., 2015; Chen et al., 2016; Zhou et al., 2018). While there were no specific genes that have known functions related to FPHE in *Salix*, the strong association for this locus also coincides with all previous reports of the SDR being located on chr15 across multiple *Salix*.

4.5.8 Pathology

Previous work using *S. viminalis* × *S. schwerinii* mapping populations (Sulima et al., 2017; Samils et al., 2011; Hanley et al., 2011; Rönnberg-Wästljung et al., 2008) have identified many QTL for leaf rust (*Melampsora* spp.) uredinia number and size, as well as leaf rust susceptibility on linkage groups representing chr01. Lacking a reference genome for *S. viminalis*, these reports were based on alignments to the *P. trichocarpa* reference genome. Here, we have identified QTL for both RUST-1 and RUST-2 in an F_2 *S. purpurea* mapping population with overlapping LOD support on *S. purpurea* chr01. Low power in the 2017 survey may be attributed to inconsistent natural inoculation of willows due to unfavorable environmental conditions or differing pathotype of rust populations between years, resulting in reduced virulence. If the latter is the case, it is possible the chr01 locus interacts with a broad range of rust pathotypes and may contain genes common to intervals reported for *S. viminalis*, like the *Salix Rust Resistance 1* locus described in Hanley et al. (2011).

While no specific candidate genes have been reported, all LOD support intervals for RUST in *S. purpurea* comprise of numerous gene models annotated as containing a nucleotide-binding site-leucine-rich repeat (NBS-LRR) motif and known pathogenesis-related genes identified in related species. Approximately 450 gene models have been annotated as containing a NBS-LRR motif in the *S. purpurea* v1.0 assembly, which is very close to the number annotated in *P. trichocarpa* (Kohler et al., 2008). Additional mapping for RUST in diverse *Salix* spp. crosses would improve our understanding of *Melampsora* pathotype diversity as well as species-specific R-gene mediated resistance mechanisms in willows.

4.6 Conclusion

The genetic dissection of complex traits in related and unrelated mapping populations can help moderate spurious associations, especially in non-model crops. This study compared association and linkage mapping methods to identify genomic regions associated with important biomass-related traits in *S. purpurea* using a naturalized association panel and full-sib F_2 family. While a number of significant marker-trait associations were identified in GWAS, only a handful of those loci were within corresponding LOD support intervals of QTL for corresponding or correlated traits in the F_2. On the converse, linkage mapping was largely successful in the F_2 family. There were many QTL in the F_2 with consistent overlapping LOD support on linkage groups between measurement years, especially for traits within QTL hotspots on *Salix* chr05 and chr10. Genetic improvement requires there be sufficient genotypic variance in a population, or progress becomes too difficult to attain. The lack of power to detect major-effect loci in the association panel was most likely a result of the small sample size, and is where progress could be made. If *S. purpurea* experienced a genetic bottleneck after naturalization, genetic improvement efforts may profit most by exploiting recent recombination events of diverse

crosses, of which, *Salix* is particularly amenable to. While linkage mapping in the F_2 produced large-effect QTL for many important biomass traits, additional replication across environments or including additional bi-parental crosses would help substantiate biomass QTL for use in downstream marker-assisted selection.

4.6 REFERENCES

Allwright MR, Taylor G. 2016. Molecular breeding for improved second generation bioenergy crops. *Trends in Plant Science,* 21: 43–54.

Aughey GN, Liu JL. 2016. Metabolic regulation via enzyme filamentation. *Critical Reviews in Biochemistry and Molecular Biology,* 51: 282–293.

Bates D, Mächler M, Bolker B, Walker S. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software,* 67: 48.

Bhandari S, Fujino T, Thammanagowda S, Zhang D, Xu F, Joshi CP. 2006. Xylem-specific and tension stress-responsive coexpression of KORRIGAN endoglucanase and three secondary wall-associated cellulose synthase genes in aspen trees. *Planta,* 224: 828–837.

Box GEP, Cox DR. 1964. An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* 26: 211–252.

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. 2007. TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics,* 23: 2633–2635.

Broman KW. 2018. qtl2: Quantitative trait locus mapping in experimental crosses.

Broman KW, Wu H, Sen S, Churchill GA. 2003. R/qtl: QTL mapping in experimental crosses. *Bioinformatics,* 19: 889–890.

Brown DM, Zeef LAH, Ellis J, Goodacre R, Turner SR. 2005. Identification of novel genes in *Arabidopsis* involved in secondary cell wall formation using expression profiling and reverse genetics. *The Plant Cell,* 17: 2281–2295.

Burke JJ, Chen J. 2015. Enhancement of reproductive heat tolerance in plants. *PLoS ONE,* 10: e0122933.

Cameron KD, Phillips IS, Kopp RF, Volk TA, Maynard CA, Abrahamson LP, Smart LB. 2008. Quantitative genetics of traits indicative of biomass production and heterosis in 34 full-sib F_1 *Salix eriocephala* families. *BioEnergy Research,* 1: 80–90.

Carlson CH, Choi Y, Chan AP, Serapiglia MJ, Town CD, Smart LB. 2017. Dominance and sexual dimorphism pervade the *Salix purpurea* L. transcriptome. *Genome Biology and Evolution*, 9: 2377–2394.

Chaves AL, Vergara CE, Mayer JE. 1995. Dichloromethane as an economic alternative to chloroform in the extraction of DNA from plant tissues. *Plant Molecular Biology Reporter*, 13: 18–25.

Che-Castaldo C, Crisafulli CM, Bishop JG, Fagan WF. 2015. What causes female bias in the secondary sex ratios of the dioecious woody shrub *Salix sitchensis* colonizing a primary successional landscape? *American Journal of Botany*, 102: 1309–1322.

Chen Y, Wang T, Fang L, Li X, Yin T. 2016. Confirmation of single-locus sex determination and female heterogamety in willow based on linkage analysis. *PLoS ONE,* 11: e0147671.

Cho J-S, Nguyen VP, Jeon H-W, Kim M-H, Eom SH, Lim YJ, Kim W-C, Park E-J, Choi Y-I, Ko J-H. 2016. Overexpression of PtrMYB119, a R2R3-MYB transcription factor from *Populus trichocarpa*, promotes anthocyanin production in hybrid poplar. *Tree Physiology*, 36: 1162–1176.

Collard B, Vera CC, McNally K, Virk P, Mackill D. 2008. Rice molecular breeding laboratories in the genomics era: Current status and future considerations. *International Journal of Plant Genomics,* 2008: 524–547.

Cronk QCB, Needham I, Rudall PJ. 2015. Evolution of catkins: Inflorescence morphology of selected Salicaceae in an evolutionary and developmental context. *Frontiers in Plant Science,* 6: 1–13.

Crossa J, Federer WT. 2012. I.4 Screening experimental designs for quantitative trait loci, association mapping, genotype-by environment interaction, and other investigations. *Frontiers in Physiology,* 3.

Davies G, Henrissat B. 1995. Structures and mechanisms of glycosyl hydrolases. *Structure,* 3: 853-859.

Denis M, Favreau B, Ueno S, Camus-Kulandaivelu L, Chaix G, Gion JM, Nourrisier-Mountou S, Polidori J, Bouvet JM. 2013. Genetic variation of wood chemical traits and association with underlying genes in *Eucalyptus urophylla*. *Tree Genetics & Genomes,* 9: 927–942.

Dixon DP, Lapthorn A, Edwards R. 2002. Plant glutathione transferases. *Genome Biology,* 3: reviews3004.1.

Dixon DP, Steel PG, Edwards R. 2011. Roles for glutathione transferases in antioxidant recycling. *Plant Signaling & Behavior,* 6: 1223–1227.

Du Q, Gong C, Wang Q, Zhou D, Yang H, Pan W, Li B, Zhang D. 2016. Genetic architecture of growth traits in *Populus* revealed by integrated quantitative trait locus (QTL) analysis and association studies. *New Phytologist,* 209: 1067–1082.

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE,* 6.

Evans LM, Slavov GT, Rodgers-Melnick E, Martin J, Ranjan P, Muchero W, Brunner AM, Schackwitz W, Gunter L, Chen J-G, Tuskan GA, DiFazio SP. 2014a. Population genomics of *Populus trichocarpa* identifies signatures of selection and adaptive trait associations. *Nature Genetics,* 46: 1089–1096.

Evans LM, Slavov GT, Rodgers-Melnick E, Martin J, Ranjan P, Muchero W, Brunner AM, Schackwitz W, Gunter L, Chen JG, Tuskan GA, DiFazio SP. 2014b. Population genomics of *Populus trichocarpa* identifies signatures of selection and adaptive trait associations. *Nature Genetics,* 46: 1089–1096.

Fahrenkrog AM, Neves LG, Resende MF, Jr., Vazquez AI, de Los Campos G, Dervinis C, Sykes R, Davis M, Davenport R, Barbazuk WB, Kirst M. 2017. Genome-wide association study reveals putative regulators of bioenergy traits in *Populus deltoides*. *New Phytologist,* 213: 799–811.

Fan M, Xu C, Xu K, Hu Y. 2012. LATERAL ORGAN BOUNDARIES DOMAIN transcription factors direct callus formation in *Arabidopsis* regeneration. *Cell Research,* 22: 1169–1180.

Fox J, Weisberg S. 2011. *An R companion to applied regression*: Sage Publications.

Furuta T, Ashikari M, Jena KK, Doi K, Reuscher S. 2017. Adapting genotyping-by-sequencing for rice F_2 populations. *G3: Genes, Genomes, Genetics,* 7: 881-893.

Gaut BS, Long AD. 2003. The lowdown on linkage disequilibrium. *The Plant Cell,* 15: 1502–1506.

Ghelardini L, Berlin S, Weih M, Lagercrantz U, Gyllenstrand N, Ronnberg-Wastljung AC. 2014. Genetic architecture of spring and autumn phenology in *Salix*. *BMC Plant Biology*, 14: 31.

Goodwin S, McPherson JD, McCombie WR. 2016. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17: 333–351.

Grattapaglia D, Sederoff R. 1994. Genetic linkage maps of *Eucalyptus grandis* and *Eucalyptus urophylla* using a pseudo-testcross: mapping strategy and RAPD markers. *Genetics*, 137: 1121–1137.

Hallauer A, Miranda J. 1988. *Quantitative genetics in plant breeding.* Ames, IA: Iowa State Univ. Press.

Hallingback HR, Fogelqvist J, Powers SJ, Turrion-Gomez J, Rossiter R, Amey J, Martin T, Weih M, Gyllenstrand N, Karp A, Lagercrantz U, Hanley SJ, Berlin S, Ronnberg-Wastljung AC. 2016. Association mapping in *Salix viminalis* L. (Salicaceae) - identification of candidate genes associated with growth and phenology. *GCB Bioenergy*, 8: 670–685.

Hanley SJ, Karp A. 2013. Genetic strategies for dissecting complex traits in biomass willows (*Salix* spp.). *Tree Physiology*, 34: 1167–80.

Hanley SJ, Pei MH, Powers SJ, Ruiz C, Mallott MD, Barker JH, Karp A. 2011. Genetic mapping of rust resistance loci in biomass willow. *Tree Genetics & Genomes*, 7: 597–608.

Hossain MA, Noh HN, Kim KI, Koh EJ, Wi SG, Bae HJ, Lee H, Hong SW. 2010. Mutation of the chitinase-like protein-encoding AtCTL2 gene enhances lignin accumulation in dark-grown *Arabidopsis* seedlings. *Journal of Plant Physiology*, 167: 650–658.

Hou J, Ye N, Zhang D, Chen Y, Fang L, Dai X, Yin T. 2015. Different autosomes evolved into sex chromosomes in the sister genera of *Salix* and *Populus*. *Scientific Reports,* 5: 1–6.

Huang X, Wei X, Sang T, Zhao Q, Feng Q, Zhao Y, Li C, Zhu C, Lu T, Zhang Z, Li M, Fan D, Guo Y, Wang A, Wang L, Deng L, Li W, Lu Y, Weng Q, Liu K, Huang T, Zhou T, Jing Y, Li W, Lin Z, Buckler ES, Qian Q, Zhang QF, Li J, Han B. 2010. Genome-wide association studies of 14 agronomic traits in rice landraces. *Nature Genetics,* 42: 961–967.

Ilegems M, Douet V, Meylan-Bettex M, Uyttewaal M, Brand L, Bowman JL, Stieger PA. 2010. Interplay of auxin, KANADI and Class III HD-ZIP transcription factors in vascular tissue formation. *Development,* 137: 975–984.

Ingvarsson PK. 2008. Multilocus patterns of nucleotide polymorphism and the demographic history of *Populus tremula*. *Genetics,* 180: 329–340.

Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, Sabatti C, Eskin E. 2010. Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics,* 42: 348–354.

Karp A, Hanley SJ, Trybush S, Macalpine W, Pei MH, Shield I. 2011. Genetic improvement of willow for bioenergy and biofuels. *Journal of Integrative Plant Biology,* 53.

Khan MA, Korban SS. 2012. Association mapping in forest trees and fruit crops. *Journal of Experimental Botany,* 63: 4045–4060.

Kim C, Guo H, Kong W, Chandnani R, Shuang LS, Paterson AH. 2016. Application of genotyping by sequencing technology to a variety of crop breeding programs. *Plant Science,* 242.

Kohler A, Rinaldi C, Duplessis S, Baucher M, Geelen D, Duchaussoy F, Meyers B, Boerjan W, Martin F. 2008. Genome-wide identification of NBS resistance genes in *Populus trichocarpa*. *Plant Molecular Biology,* 66: 619–36.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics,* 25: 1754–60.

Linster CL, Clarke SG. 2008. L-Ascorbate biosynthesis in higher plants: The role of VTC2. *Trends in Plant Science,* 13: 567–573.

Lipka AE, Tian F, Wang Q, Peiffer J, Li M, Bradbury PJ, Gore MA, Buckler ES, Zhang Z. 2012. GAPIT: genome association and prediction integrated tool. *Bioinformatics,* 28: 2397–2399.

Mackay TF, Stone EA, Ayroles JF. 2009. The genetics of quantitative traits: Challenges and prospects. *Nature Reviews Genetics,* 10: 565–577.

Majer C, Hochholdinger F. 2011. Defining the boundaries: Structure and function of LOB domain proteins. *Trends in Plant Science,* 16: 47–52.

Marino D, Dunand C, Puppo A, Pauly N. 2012. A burst of plant NADPH oxidases. *Trends in Plant Science,* 17: 9–15.

McKown AD, Klápště J, Guy RD, Geraldes A, Porth I, Hannemann J, Friedmann M, Muchero W, Tuskan GA, Ehlting J, Cronk QCB, El-Kassaby YA, Mansfield SD, Douglas CJ. 2014. Genome-wide association implicates numerous genes underlying ecological trait variation in natural populations of *Populus trichocarpa*. *New Phytologist,* 203: 535–553.

Merret R, Carpentier MC, Favory JJ, Picart C, Descombin J, Bousquet-Antonelli C, Tillard P, Lejay L, Deragon JM, Charng YY. 2017. Heat shock protein HSP101 affects the release of ribosomal protein mRNAs for recovery after heat shock. *Plant Physiology,* 174: 1216–1225.

Meuwissen TH, Hayes BJ, Goddard ME. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics,* 157: 1819–29.

Mhamdi A, Mauve C, Gouia H, Saindrenan P, Hodges M, Noctor G. 2010. Cytosolic NADP-dependent isocitrate dehydrogenase contributes to redox homeostasis and the regulation of pathogen responses in *Arabidopsis* leaves. *Plant, Cell & Environment,* 33: 1112–1123.

Money D, Gardner K, Migicovsky Z, Schwaninger H, Zhong GY, Myles S. 2015. LinkImpute: Fast and accurate genotype imputation for nonmodel organisms. *G3: Genes, Genomes, Genetics,* 5: 2383–2390.

Myles S, Boyko AR, Owens CL, Brown PJ, Grassi F, Aradhya MK, Prins B, Reynolds A, Chia JM, Ware D, Bustamante CD, Buckler ES. 2011. Genetic structure and domestication history of the grape. *Proceedings of the National Academy of Sciences of the United States of America,* 108: 3530–3535.

Olson MS, Levsen N, Soolanayakanahally RY, Guy RD, Schroeder WR, Keller SR, Tiffin P. 2013. The adaptive potential of *Populus balsamifera* L. to phenology requirements in a warmer global climate. *Molecular Ecology,* 22: 1214–1230.

Olson MS, Robertson AL, Takebayashi N, Silim S, Schroeder WR, Tiffin P. 2010. Nucleotide diversity and linkage disequilibrium in balsam poplar (*Populus balsamifera*). *New Phytologist,* 186: 526–536.

Payne CT, Zhang F, Lloyd AM. 2000. GL3 encodes a bHLH protein that regulates trichome development in *Arabidopsis* through interaction with GL1 and TTG1. *Genetics,* 156: 1349–1362.

Pose D, Castanedo I, Borsani O, Nieto B, Rosado A, Taconnat L, Ferrer A, Dolan L, Valpuesta V, Botella MA. 2009. Identification of the *Arabidopsis* dry2/sqe1-5 mutant reveals a central role for sterols in drought tolerance and regulation of reactive oxygen species. *Plant Journal,* 59: 63–76.

Preston JC, Hileman LC. 2013. Functional evolution in the plant SQUAMOSA-PROMOTER BINDING PROTEIN-LIKE (SPL) gene family. *Frontiers in Plant Science,* 4: 80.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics,* 155: 945–959.

Prunier J, Pelgas B, Gagnon F, Desponts M, Isabel N, Beaulieu J, Bousquet J. 2013. The genomic architecture and association genetics of adaptive characters using a candidate SNP approach in boreal black spruce. *BMC Genomics,* 14: 368.

Pucholt P, Rönnberg-Wästljung A-C, Berlin S. 2015. Single locus sex determination and female heterogamety in the basket willow (*Salix viminalis* L.). *Heredity,* 114: 575–583.

R Core Team. 2015. *R: A language and environment for statistical computing.* Vienna, Australia: R Foundation for Statistical Computing.

Rae AM, Street NR, Robinson KM, Harris N, Taylor G. 2009. Five QTL hotspots for yield in short rotation coppice bioenergy poplar: The Poplar Biomass Loci. *BMC Plant Biology,* 9: 23.

Rasbery JM, Shan H, LeClair RJ, Norman M, Matsuda SPT, Bartel B. 2007. *Arabidopsis thaliana* squalene epoxidase 1 is essential for root and seed development. *Journal of Biological Chemistry*, 282: 17002–17013.

Rodríguez-Álvarez MX, Boer M, Eilers P. 2017. SpATS: Spatial analysis of field trials with splines. *Theoretical and Applied Genetics*, 130: 1375–1392.

Rodríguez-Álvarez MX, Lee D-J, Kneib T, Durbán M, Eilers P. 2015. Fast smoothng parameter separaton n multdmensonal generalzed P-splnes: the SAP algorthm. *Statistics and Computing*, 25: 941–957.

Rönnberg-Wästljung A-C, Samils B, Tsarouhas V, Gullberg U. 2008. Resistance to *Melampsora larici-epitea* leaf rust in *Salix*: Analyses of quantitative trait loci. *Journal of Applied Genetics*, 49: 321–331.

Rosti J, Barton CJ, Albrecht S, Dupree P, Pauly M, Findlay K, Roberts K, Seifert GJ. 2007. UDP-glucose 4-epimerase isoforms UGE2 and UGE4 cooperate in providing UDP-galactose for cell wall biosynthesis and growth of *Arabidopsis thaliana*. *Plant Cell*, 19: 1565–1579.

Samils B, Rönnberg-Wästljung A-C, Stenlid J. 2011. QTL mapping of resistance to leaf rust in *Salix*. *Tree Genetics & Genomes*, 7: 1219–1235.

Sannigrahi P, Ragauskas AJ, Tuskan GA. 2010. Poplar as a feedstock for biofuels: A review of compositional characteristics. *Biofuels, Bioproducts and Biorefining*, 4: 209–226.

Saska MM, Kuzovkina YA. 2010. Phenological stages of willow (*Salix*). *Annals of Applied Biology*, 156: 431–437.

Serapiglia M, Cameron K, Stipanovic A, Smart L. 2009. Analysis of biomass composition using high-resolution thermogravimetric analysis and percent bark content for the selection of shrub willow bioenergy crop varieties. *BioEnergy Research,* 2: 1–9.

Serapiglia MJ, Cameron KD, Stipanovic AJ, Smart LB. 2012. Correlations of expression of cell wall biosynthesis genes with variation in biomass composition in shrub willow (*Salix* spp.) biomass crops. *Tree Genetics & Genomes,* 8: 775–788.

Serapiglia MJ, Gouker FE, Smart LB. 2014. Early selection of novel triploid hybrids of shrub willow with improved biomass yield relative to diploids. *BMC Plant Biology,* 14: 74.

Shapiro SS, Wilk MB. 1965. An analysis of variance test for normality (complete samples). *Biometrika,* 52: 591–611.

Shikata M, Matsuda Y, Ando K, Nishii A, Takemura M, Yokota A, Kohchi T. 2004. Characterization of Arabidopsis ZIM, a member of a novel plant-specific GATA factor gene family. *Journal of Experimental Botany,* 55: 631–639.

Slavov GT, DiFazio SP, Martin J, Schackwitz W, Muchero W, Rodgers-Melnick E, Lipphardt MF, Pennacchio CP, Hellsten U, Pennacchio LA, Gunter LE, Ranjan P, Vining K, Pomraning KR, Wilhelm LJ, Pellegrini M, Mockler TC, Freitag M, Geraldes A, El-Kassaby YA, Mansfield SD, Cronk QC, Douglas CJ, Strauss SH, Rokhsar D, Tuskan GA. 2012. Genome resequencing reveals multiscale geographic structure and extensive linkage disequilibrium in the forest tree *Populus trichocarpa*. *New Phytologist,* 196: 713–725.

Smirnoff N, Wheeler GL. 2000. Ascorbic acid in plants: Biosynthesis and function. *Critical Reviews in Biochemistry and Molecular Biology,* 35: 291–314.

Soto-Cerda BJ, Duguid S, Booker H, Rowland G, Diederichsen A, Cloutier S. 2014. Genomic regions underlying agronomic traits in linseed (*Linum usitatissimum* L.) as revealed by association mapping. *Journal of Integrative Plant Biology,* 56: 75–87.

Sulima P, Przyborowski JA, Kuszewska A, Załuski D, Jędryczka M, Irzykowski W. 2017. Identification of quantitative trait loci conditioning the main biomass yield components and resistance to *Melampsora* spp. in *Salix viminalis* × *Salix schwerinii* hybrids. *International Journal of Molecular Sciences,* 18: 677.

Tang Y, Liu X, Wang J, Li M, Wang Q, Tian F, Su Z, Pan Y, Liu D, Lipka AE, Buckler ES, Zhang Z. 2016. GAPIT version 2: An enhanced integrated tool for genomic association and prediction. *Plant Genome,* 9.

Taylor J, Butler D. 2017. R package ASMap: Efficient genetic linkage map construction and diagnosis. *Journal of Statistical Software,* 79: 1–29.

Thatcher LF, Kazan K, Manners JM. 2012. Lateral organ boundaries domain transcription factors: New roles in plant defense. *Plant Signaling & Behavior,* 7: 1702–1704.

Urbanowicz BR, Bennett AB, del Campillo E, Catalá C, Hayashi T, Henrissat B, Höfte H, McQueen-Mason SJ, Patterson SE, Shoseyov O. 2007. Structural organization and a standardized nomenclature for plant endo-1, 4-β-glucanases (cellulases) of glycosyl hydrolase family 9. *Plant Physiology,* 144: 1693–1696.

Valentine J, Clifton-Brown J, Hastings A, Robson P, Allison G, Smith P. 2012. Food vs. fuel: The use of land for lignocellulosic 'next generation' energy crops that minimize competition with primary food production. *GCB Bioenergy,* 4: 1–19.

Velazco JG, Rodriguez-Alvarez MX, Boer MP, Jordan DR, Eilers PHC, Malosetti M, van Eeuwijk FA. 2017. Modelling spatial trends in sorghum breeding field trials using a two-dimensional P-spline mixed model. *Theoretical and Applied Genetics*, 130: 1375–1392.

Wang Q, Tian F, Pan Y, Buckler ES, Zhang Z. 2014. A SUPER powerful method for genome wide association study. *PLoS ONE*, 9: e107684.

Wu G, Lin W-C, Huang T, Poethig RS, Springer PS, Kerstetter RA. 2008a. KANADI1 regulates adaxial–abaxial polarity in *Arabidopsis* by directly repressing the transcription of ASYMMETRIC LEAVES2. *Proceedings of the National Academy of Sciences of the United States of America*, 105: 16392–16397.

Wu Y, Bhat PR, Close TJ, Lonardi S. 2008b. Efficient and accurate construction of genetic linkage maps from the minimum spanning tree of a graph. *PLoS Genetics,* 4: e1000212.

Yan JB, Shah T, Warburton ML, Buckler ES, McMullen MD, Crouch J. 2009. Genetic characterization and linkage disequilibrium estimation of a global maize collection using SNP markers. *PLoS ONE*, 4: e8451.

Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL. 2014. Advantages and pitfalls in the application of mixed-model association methods. *Nature Genetics*, 46: 100–106.

Zhang F, Gonzalez A, Zhao M, Payne CT, Lloyd A. 2003. A network of redundant bHLH proteins functions in all TTG1-dependent pathways of *Arabidopsis*. *Development,* 130: 4859–4869.

Zhang Z, Ersoz E, Lai CQ, Todhunter RJ, Tiwari HK, Gore MA, Bradbury PJ, Yu J, Arnett DK, Ordovas JM, Buckler ES. 2010. Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics*, 42: 355–360.

Zhou R, Macaya-Sanz D, Rodgers-Melnick E, Carlson CH, Gouker FE, Evens LM, Schmutz J, Jenkins JW, Yan J, Tuskan GA, Smart LB, DiFazio SP. 2018. Characterization of a large sex determination region in *Salix purpurea* L. (Salicaceae). *Molecular Genetics and Genomics*, DOI: 10.1007/s00438-018-1473-y.

CHAPTER 5

DOMINANCE AND SEXUAL DIMORPHISM PERVADE THE *SALIX PURPUREA* L. TRANSCRIPTOME

Published as: Carlson CH, Choi Y, Chan AP, Town CD, Smart LB. 2017. *Genome Biology and Evolution*, 9(9): 2377–2394.

5.1 Abstract

The heritability of gene expression is critical in understanding heterosis and is dependent on allele-specific regulation by local and remote factors in the genome. We used RNA-Seq to test whether variation in gene expression among F_1 and F_2 intraspecific *Salix purpurea* progeny is attributable to *cis*- and *trans*-regulatory divergence. We assessed the mode of inheritance based on gene expression levels and allele-specific expression for F_1 and F_2 intraspecific progeny in two distinct tissue types: shoot tip and stem internode. In addition, we explored sexually dimorphic patterns of inheritance and regulatory divergence among F_1 progeny individuals. We show that in *S. purpurea* intraspecific crosses, gene expression inheritance largely exhibits a maternal dominant pattern, regardless of tissue type or pedigree. A significantly greater number of *cis*- and *trans*-regulated genes coincided with upregulation of the maternal parent allele in the progeny, irrespective of the magnitude, whereas the paternal allele was higher expressed for genes showing *cis* × *trans* or compensatory regulation. Importantly, consistent with previous genetic mapping results for sex in shrub willow, we have delimited sex-biased gene expression to a 2 Mb pericentromeric region on *S. purpurea* chr15 and further refined the sex determination region. Altogether, our results offer insight into the inheritance of gene expression in *S. purpurea*

as well as evidence of sexually dimorphic expression which may have contributed to the evolution of dioecy in *Salix*.

5.2 Introduction

Allele-specific expression (ASE) reflects the regulatory status of each parent allele inherited in an individual and has become an informative phenotype for biologists in understanding nonadditive phenotypic expression (Stupar and Springer, 2006). Without further knowledge of parent pedigree, ASE can only be considered for sites that differ between the parents of an F_1 cross, whereby a single copy of each homozygous parent allele exists in a heterozygous state in the F_1 hybrid. For any biallelic site, the normalized expression ratio of the female parent (P1) allele and the male parent (P2) allele is contrasted to the same ASE ratio (P1H/P2H) in the hybrid. Statistically significant deviations of ASE in the F_1 from the expected contribution of each parent (Pr = 0.5) are based on binomial exact tests, which lay the groundwork for the estimation of *cis*-regulatory divergent gene expression. Despite the fact that the *extent* of regulatory divergence is largely dependent on pre-determined global significance thresholds (i.e., False Discovery Rate, FDR), the overall *patterns* of divergent expression do not drastically change (Suvorov et al., 2013). These patterns are broadly subject to sequence variation observed in the domains of local *cis*-regulatory elements or remote *trans*-acting factors.

There is evidence that nonadditive gene expression can confer novel transgressive phenotypes in hybrids (Springer and Stupar, 2007) and is alleged to be a major driver of hybrid speciation (Rieseberg et al., 2003). For instance, in interspecific crosses of *Drosophilia melanogaster* and *D. simulans*, *cis*-effects were shown to account for a majority of regulatory divergent expression (Wittkopp et al., 2008a), whereas *trans*-effects accounted for a higher

proportion of expression variation between parents of the same species (Wittkopp et al., 2004; Wittkopp et al., 2008b). Hybridization can introduce substantial divergence in offspring gene expression when compared with that of the parents (McManus et al., 2010). Such a merger provides new allelic variation within the regulatory domains of genes (e.g., promoters) as well as new targets of *trans*-acting factors (e.g., transcription factors). While mutations in *cis*-regulatory elements have been shown to account for evolutionarily significant phenotypic change, *trans*-regulatory evolution can also affect adaptive morphological change (Wittkopp and Kalay, 2011).

Depending on the effective population size, the effects of *cis*-mutations on gene expression are generally considered to be less deleterious as they only affect a single gene and are more likely to become fixed, whereas *trans*-effects can alter the expression of a number of genes (pleiotropy) and are more likely to be subject to purifying selection (McManus et al., 2010). Consequently, the conservation of gene expression ($P1 \approx H \approx P2$) should be less pronounced in wide hybrids, but more common with inbreeding or sib-mating, because the transcriptional activity of ASE is simply a function of the two *cis*-regulatory parent alleles in a common *trans*-regulated background. Nevertheless, studies of ASE in intraspecific progeny derived from closely-related parents have attributed parental expression divergence to both *cis*- and *trans*- regulatory components (Bell et al., 2013; Suvorov et al., 2013; Combes et al., 2015).

A bulk of ASE work in plants have used hybrids derived from inbred parents to study the effects of hybridization on gene expression (Stupar and Springer, 2006; Song et al., 2013). Although there are notable exceptions (Bell et al., 2013; Comai et al., 2015), there is a general lack of understanding on the evolution of gene expression with regards to the hybridization of heterozygous parents from natural, obligate outcrossing populations. Previous expression studies in dioecious shrub willow (*Salix* spp., Salicaceae) have predominantly focused on correlating

functional variation of candidate gene family members to lignocellulosic composition traits in contrasting pedigrees (Puckett et al., 2012; Serapiglia et al., 2012). In this study, I examined the variation in transcriptome-wide expression within and among full-sib F_1 and F_2 intraspecific families generated from heterozygous parents collected from naturalized *S. purpurea* L. populations.

Shrub willow has been bred as a dedicated energy crop since the early 1970's with the goal of producing fast-growing bioenergy feedstock cultivars that are high-yielding, genetically diverse, pest and disease resistant, and able to grow on marginal land without competing with food crops (Stoof et al., 2014). The heterogeneity and adaptive plasticity of *Salix* spp. provides an abundant germplasm pool for trait improvement and phylogenetic characterization (Hanley and Karp, 2013). Hybridization is a key component in the development of shrub willow bioenergy crops, as hybrids often display heterosis for yield (Fabio et al., 2016). While significant improvements in biomass yield has been realized in interspecific crosses of *Salix* (Kopp et al., 2001; Cameron et al., 2008), heterosis is more pronounced in triploid progeny derived from the hybridization of diploid and tetraploid parents (Smart and Cameron, 2008; Serapiglia et al., 2014a; Carlson and Smart, 2016). These high-yielding triploid shrub willow outperform foundation commercial cultivars and show promise for the future of the biomass production industry (Serapiglia et al., 2014b; Fabio et al., 2017).

With the public release of the *S. purpurea* genome reference assembly (phytozome.jgi.doe.gov), *Salix* has become a powerful model to study the genomic basis of heterosis in dioecious species. *Salix purpurea* has a relatively compact genome (~400 Mb) with ~37,500 primary gene models and ~65,000 alternatively-spliced isoforms (Smart et al., in prep). Although the genome of *S. purpurea* is remarkably collinear to that of *Populus trichocarpa*

(Berlin et al., 2010), major differences in the overall arrangement and abundance of coding and non-coding DNA (Hou et al., 2016) has radically affected the ecology, habit, and reproductive paths since the Salicoid duplication and divergence of the genera (Rodgers-Melnick et al., 2012). For instance, the sex determining region (SDR) of *P. trichocarpa* resides within a peritelomeric region on *Populus* chr19 (Yin et al., 2008; McKown et al., 2017), whereas the SDR of *Salix* spp. has been mapped to a pericentromeric region on *Salix* chr15 (Pucholt et al., 2015; Zhou et al., 2017). In addition to contrasting genomic locations of the SDR, *S. purpurea* exhibits a ZW sex determining system with heterogametic females (Zhou et al., 2018) and *P. trichocarpa*, an XY system with heterogametic males (Tuskan et al., 2012). To date, the mechanism of sex determination in the Salicaceae has not been completely resolved.

The main objectives of this study were (1) to test for differential gene expression among the shoot tip and internode transcriptome of segregating F_1 and F_2 intraspecific *S. purpurea* family progeny, (2) to categorize gene expression by modes of inheritance, (3) to assess the magnitude and direction of regulatory divergent expression, and (4) to examine the regulatory components of sexually dimorphic gene expression that may have contributed to the evolution of dioecy in *Salix*.

5.3 Materials and Methods

5.3.1 Plant Material and Growing Conditions

The full-sib intraspecific F_1 *S. purpurea* family was generated from a cross between the female clone 94006 and the male clone 94001, collected from a population of *S. purpurea* in Central New York. Two F_1 siblings from this family were selected and crossed (9882-41 × 9882-34) to generate the F_2 population (Figure 5.1). Dormant first-year post-coppice vegetative shoots of all family parents and progeny were taken from nursery beds in the winter of 2013. Cuttings

of equal length (20 cm) and diameter were cut from shoots and planted into 2.5 L containers filled with Farfard® PV-1 potting media and grown under environmentally-controlled greenhouse conditions with supplemental lighting provided on a 14h day : 10 h night regimen with a max daytime temperature of 26°C and a nighttime temperature of 18°C. All plots were completely randomized over five replicate blocks. Liquid fertilizer (Peter's 15-16-17 Peat-Lite Special; Scotts Miracle-Gro Company, Marysville, Ohio, USA) was applied weekly at 100 ppm after the third week from planting cuttings, until the study was terminated. Herein, I refer to parents by their clone identifiers and discriminate the female and male parents as P1 and P2, respectively.

5.3.2 Determination of Ploidy Level

The relative DNA content (pg 2C value^{-1}) of family parents and progeny was determined by flow cytometric analysis using young leaf material harvested from actively growing shoots in greenhouse conditions. Analysis of 50 mg of mature leaf tissue from parental genotypes and selected progeny was performed at the Flow Cytometry and Imaging Core Laboratory at Virginia Mason Research Center in Seattle, WA as was previously described (Serapiglia et al., 2014b). A minimum of four replicates of all samples were independently assessed using either the diploid *S. purpurea* female genome reference clone 94006 or the diploid *S. purpurea* male clone 94001 as an internal standard. Diploid parent clones from multiple runs were averaged and then divided by the value of the check for that run. When a clone was analyzed more than once, 2C values were averaged. All parents and progeny described in this study are diploid (2n=38) according to flow cytometric and genetic marker analysis (Argus 1997; Serapiglia et al., 2014a).

5.3.3 RNA Sample Preparation and Sequencing

Total RNA was extracted from three biological replicates of 10 random family progeny and their parents for the F_1 and F_2 families. Both shoot tip and internode tissues from each

individual were collected and processed using the Spectrum™ Total Plant RNA Kit with DNase digestion (Sigma, St. Louis, MO), following the manufacturers procedures. Cold-ethanol precipitations were performed by addition of 10 µl actetic acid and 280 µl 100% cold ethanol to 100 µl eluate and placed in -80°C for at least 3 hr. Samples were centrifuged at 17,000 × *g* for 30 min at 4°C, washed with 80% cold ethanol, then centrifuged at 17,000 × *g* for 20 min at 4°C. After centrifugation, the ethanol supernatant was discarded and the pellet re-suspended in ribonuclease-free 10 mM Tris-HCl (pH = 8). Quantification of sample quality and concentration was performed using the Experion RNA StdSens kit (Bio-Rad Laboratories, Inc., Hercules, CA), following manufacturers' procedures.

Independent extractions were performed on three replicate plants of each of 10 progeny individuals within each family and subsequently pooled in equal concentrations. For each tissue type, three RNA-Seq libraries were constructed representing the female parent, the male parent, and a pool of of 10 progeny. In addition, for comparisons between progeny within the F_1 family, 10 F_1 progeny were individually barcoded and sequenced. Libraries were constructed using the NEBNext Ultra Directional RNA Library Prep Kit and sequenced on the Illumina platform (1x100 bp) at the J. Craig Venter Institute (Rockville, MD).

Shoot tips were defined as the shoot axis that is the most distal part of a shoot system, comprised of a shoot apical meristem and the youngest leaf primordia. Stem internodes were defined as the cardinal organ part that is the part of a shoot axis between two nodes of the axis.

5.3.4 Read Filtering, Alignment, and Variant Discovery

Low-coverage paired-end genomic DNA sequencing of the parental lines for the F_1 and F_2 families was performed to validate variants from RNA-Seq data. Biallelic SNPs were used to quantify allele-specific expression and regulatory divergence within and among intraspecific

family progeny. Parent DNA libraries were sequenced (Illumina HiSeq 2×101) and aligned to the *S. purpurea* v1 reference genome using BWA-MEM (-M -R) (Li and Durbin 2009). Subsequent SAM files were sorted, marked for duplicates, and indexed in Picard (broadinstitute.github.io/picard). Indel realignment and variant calling was performed using *HaplotypeCaller* (emit_conf=10, call_conf=30) in the Genome Analysis Toolkit (GATK) (DePristo et al., 2011). For all samples, RNA-Seq reads were trimmed (min length=50) and mapped (length=0.8, similarity=0.9) to the *S. purpurea* v1 reference genome using CLC Genomics Workbench (www.qiagenbioinformatics.com). The *S. purpurea* L. v1.0 genome reference assembly and annotation is available online via the Joint Genome Institute Comparative Plant Genomics Portal, Phytozome v12 (phytozome.jgi.doe.gov).

5.3.5 Gene Expression Inheritance Classifications

To determine the mode of inheritance for genes, the number of RNA-Seq reads mapped to individual genes was counted for each of the female (P1) and male (P2) parents and progeny (H). Expression levels were compared based on normalized read counts using the edgeR package (Robinson et al., 2010). Differentially expressed genes were determined using an exact test implemented in edgeR for negative-binomially distributed counts (disp=0.1, FDR=0.005). We used a custom R script to sort genes into the following six inheritance categories: (1) P1-dominant: H≈P1 and H≠P2, (2) P2-dominant: H≈P2 and H≠P1, (3) additive: P1<H<P2 or P2<H<P1, (4) overdominant: H>P1 and H>P2, (5) underdominant: H<P1 and H<P2, and (6) conserved: all others.

The absolute magnitude of dimorphic gene expression inheritance was determined as the Euclidean distance (L^2) between vectors m_{xy} and f_{xy}, such that: $L^2 = \langle p, q \rangle = [(|m_x| - |f_x|)^2 + (|m_y| - |f_y|)^2]^{0.5}$, where m_x is the male coordinate derived from $\log_2(m/p_1)$, m_y is

the male coordinate derived from $\log_2(m/p_2)$, f_x is the female coordinate derived from $\log_2(f/p_1)$, and f_y is the female coordinate derived from $\log_2(f/p_2)$, p is the squared absolute difference of the vectors $(|m_x|, |f_x|)$ and $(|m_x|, |m_y|)$, and q is the squared absolute difference of the vectors $(|m_x|, |f_y|)$ and $(|m_x|, |m_y|)$.

5.3.6 Regulatory Divergence Classifications

For regulatory divergence classification of genes, the sequence reads in progeny need to be assigned to their parental origins. For each gene, the expression levels of the two parental alleles were estimated based on the nucleotide allele counts across all SNP sites detected based on parent DNA libraries (described above), and where the nucleotide alleles present in each of the parents are distinct and therefore allow unambiguous assignment of parental origins. Categories of regulatory functions considered conserved, compensatory, ambiguous, *cis*, *trans*, *cis* + *trans*, and *cis* × *trans*, were sorted using R scripts, as described in Landry et al. (2005) and McManus et al. (2010). Regulatory divergence assignments were based on two sets of tests: (1) a binomial exact test between P1 and P2 in the parents and between P1H and P2H in the progeny, and (2) Fisher's Exact Test on P1, P2, P1H, and P2H (Table 5.1).

Table 5.1 Regulatory classifications.

Classification	**P1 : P2**	**P1H : P2H**	**P1 : P2 : P1H : P2H**
cis only	yes	yes	no
trans only	yes	no	yes
cis + *trans* or *cis* × *trans*	yes	yes	yes
compensatory	no	yes	yes
conserved	no	no	no
ambiguous	*all others*		

For instance, *trans* regulation was considered as significant differential expression in the parent but not in the progeny. *Cis* regulation was considered as showing a significant differential expression in the progeny and parents and no significant *trans*-effects. *Cis* + *trans* was

considered as significant differential expression in progeny and parent, and also with significant trans-effects (i.e., the parent with higher expression contributes the higher-expressed allele in progeny). *Cis* × *trans* was considered as significant differential expression in progeny and parent with significant *trans*-effects (i.e., the parent with higher expression contributes the lower-expressed allele in progeny). Conserved genes were considered as having no significant differential expression in the progeny or parents. Compensatory was defined as significant differential expression in the progeny but not in the parents but significant *trans*-effects, and genes identified as ambiguous could not be sorted into a functional category.

The percent divergence due to *cis*- and *trans*-effects were calculated such that % *cis* = [|*cis*| / (|*cis*| + |*trans*|)] × 100 and % *trans* = [|*trans*| / (|*cis*| + |*trans*|)] × 100, where *cis* = $\log_2$(P1H / P2H) and *trans* = $\log_2$(P1H / P2H) - $\log_2$(P1 / P2).

5.3.7 Tests for Differential Expression

All statistical analyses were performed in the open-source statistical computing environment, R (Team 2015). Tests for differential expression were conducted in the package edgeR (Robinson et al., 2010). Normalization factors and dispersion estimates (robust=T) were computed prior to tests for differential expression. A general linear model was used to fit normalized count data using *glmFit* and *glmLRT* to conduct likelihood ratio tests for the model coefficients. Tests for paired comparisons were conducted to investigate the effect of tissue type over the individuals within the F_1 family, using an additive linear model with clone as a blocking factor. In order to test for differential expression by sex in the F_1, a conservative and robust quasi-likelihood model using tag-wise dispersion estimates was used to fit the data with *glmQLFit* (robust=T). For tests of differential expression, genes were only considered to be significant at a False Discovery Rate of 0.05. To explore variation in gene content, I considered

the total number of genes expressed per sample library, without inferring relative expression per gene copy or per cell, as is described in Coate and Doyle (2010). A summary of analyses and corresponding libraries used is in Table 5.2.

5.3.8 Gene Ontology

Gene ontology (GO) enrichment was performed in agriGO (Du et al., 2010) using the subset of the *S. purpurea* v1 transcriptome (reference set) that passed filtering prior to tests of differential expression. Only significant ontologies were reported from query lists. For *S. purpurea* gene annotations which encode for hypothetical proteins, gene models and associated GO-terms were inferred using the best-hit (BLASTP E-value ≤ 0.1) to *Populus trichocarpa* (Phytozome v10.3 annotation) and Arabidopsis (TAIR10 and Araport11 annotations) proteome.

Table 5.2 Analyses conducted and libraries used for corresponding analyses.

Analyses	Libraries	Number of Libraries
Gene expression inheritance		
Shoot tip tissue		
F_1 family	pooled F_1 progeny + F_1 P1 + F_1 P2	3
F_2 family	pooled F_2 progeny + F_2 P1 + F_2 P2	3
Internode tissue		
F_1 family	pooled F_1 progeny + F_1 P1 + F_1 P2	3
F_2 family	pooled F_2 progeny + F_2 P1 + F_2 P2	3
Regulatory divergence		
Shoot tip tissue		
F_1 family	pooled F_1 progeny + F_1 P1 + F_1 P2	3
F_2 family	pooled F_2 progeny + F_2 P1 + F_2 P2	3
Internode tissue		
F_1 family	pooled F_1 progeny + F_1 P1 + F_1 P2	3
F_2 family	pooled F_2 progeny + F_2 P1 + F_2 P2	3
Differential expression in the F_1		
Tissue type	unpooled F_1 shoot tip, unpooled F_1 internode	20
Midparent	unpooled F_1, F_1 P1 + F_1 P2	24
Sex	unpooled F_1 females + F_1 P1, unpooled F_1 males + F_1 P2	8
Sexual dimorphism in the F_1		
Gene expression inheritance		
F_1 males	3 unpooled F_1 males + F_1 P1 + F_1 P2	5
F_1 females	3 unpooled F_1 females + F_1 P1 + F_1 P2	5
Regulatory divergence		
F_1 males	3 unpooled F_1 males + F_1 P1 + F_1 P2	5
F_1 females	3 unpooled F_1 females + F_1 P1 + F_1 P2	5

5.4 Results

5.4.1 Transcriptome Analysis

In order to define the factors contributing to variation in global expression among the F_1 and F_2 parents and progeny the RNA-Seq data were subjected to multidimensional scaling (MDS) analysis, considering only genes with a sum cpm-normalized expression >1.0 for ≥50% of the samples (Figure 5.1). As expected, the parent transcriptomes of the F_1 were the most distantly clustered, because they are the least related in this study. The first MDS dimension clustered samples based entirely on tissue type, and the second dimension split F_1 and F_2 parents by sex, then by pedigree. While 9882-41 and 9882-34 are F_1 siblings, their gene expression levels clustered more towards the parent of the same sex. The pedigrees, ploidy levels, and read mapping statistics are summarized in Table 5.3.

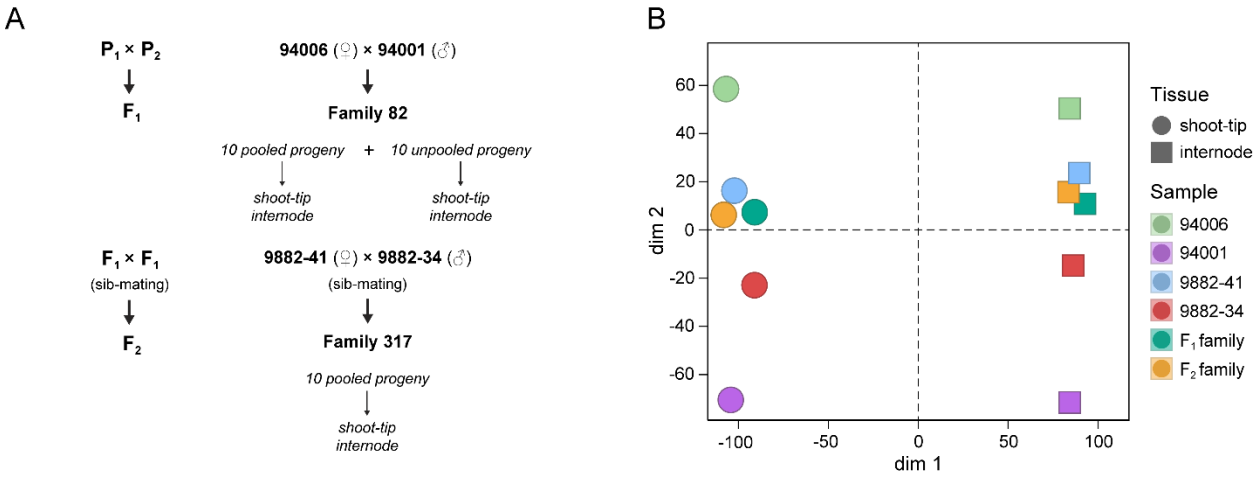

Figure 5.1. Summary of **(A)** family pedigree and experimental design and a **(B)** multidimensional scaling plot of F_1 and F_2 parents and progeny. Colored according to their pedigree, circles represent the library-normalized expression of the sample shoot tip transcriptome, and squares, the internode transcriptome.

Table 5.3 Summary of intraspecific F_1 and F_2 *Salix purpurea* sample pedigree, ploidy-level, the total number of mapped reads to the *S. purpurea* v1 reference genome, and the total number of genes expressed (cpm > 1).

Pedigree			Ploidy		Mapped Reads (1×10^6)		Genes Expressed	
Sample ID	**Generation**	**Cross (P1 × P2)**	**2N=**	**pg $2C^{-1}$ ± SD**	**Shoot-tip**	**Internode**	**Shoot-tip**	**Internode**
Parents of F_1 and F_2 families								
94006	P1 F1	natural accession	2X=38	0.886 ± 0.041	28.002	26.424	27,019	27,239
94001	P2 F1	natural accession	2X=38	1.030 ± 0.134	29.266	29.582	26,900	26,970
9882-41	P1 F2	94006 × 94001	2X=38	0.904 ± 0.030	27.011	20.484	26,708	26,723
9882-34	P2 F2	94006 × 94001	2X=38	0.963 ± 0.008	25.485	27.35	26,720	26,572
Pooled F_1 and F_2 progeny								
Family 82	F1 family	94006 × 94001	2X=38	0.898 ± 0.023	32.493	22.489	26,897	27,381
Family 317	F2 family	9882-41 × 9882-34	2X=38	0.895 ± 0.001	27.46	28.565	26,795	26,947
F_1 progeny individuals								
10X-082-013	F1 family	94006 × 94001	2X=38	0.871 ± 0.013	7.309	25.693	25,343	26,082
10X-082-018	F1 family	94006 × 94001	2X=38	0.865 ± 0.008	15.258	14.53	25,773	25,919
10X-082-030	F1 family	94006 × 94001	2X=38	0.902 ± 0.006	14.542	15.963	26,014	25,039
10X-082-035	F1 family	94006 × 94001	2X=38	0.894 ± 0.005	31.682	14.221	25,474	25,264
10X-082-060	F1 family	94006 × 94001	2X=38	0.924 ± 0.009	13.251	12.446	26,226	26,172
10X-082-062	F1 family	94006 × 94001	2X=38	0.934 ± 0.004	17.325	18.143	25,995	25,756
10X-082-067	F1 family	94006 × 94001	2X=38	0.919 ± 0.012	13.799	19.233	26,013	25,862
10X-082-071	F1 family	94006 × 94001	2X=38	0.927 ± 0.017	17.253	14.917	25,498	25,951
10X-082-078	F1 family	94006 × 94001	2X=38	0.998 ± 0.016	25.608	32.089	25,519	25,806
10X-082-093	F1 family	94006 × 94001	2X=38	0.990 ± 0.018	18.04	16.911	25,877	25,412

5.4.2 Inheritance of Gene Expression

The mode of gene expression inheritance assessed in F_1 and F_2 intraspecific progeny was based on two distinct tissue types: shoot tip and stem internode. Differentially expressed genes were defined as having normalized expression levels significantly higher or lower for treatment comparisons at an FDR of 0.005. In general, gene expression in both the F_1 and F_2 family was largely conserved, yet expression differences between the F_1 parents was far more pronounced than among the F_1 progeny and the F_1 midparent. Likewise, the parents of the F_2 family had dramatically fewer differentially expressed genes than among the parents of the F_1, whereas gene expression in the F_2 was considerably more conserved than in the F_1. Regardless of the tissue type, less than 0.07% of all differentially expressed genes in the F_1 and F_2 could be classified as having an additive mode of expression inheritance. Among differentially expressed genes, the greatest proportion were classified as dominant (Figure 5.2); accounting for 95% and 84% of all nonadditive (i.e., dominant, overdominant, and underdominant) F_1 gene expression and 94% and 96% F_2 gene expression in the internode and shoot tip transcriptome, respectively (Table 5.4).

Further, both the F_1 and F_2 families showed a significantly greater proportion of dominant expression biased in the direction of the maternal parent (P1 dominant). The most extreme case of maternal dominance was identified in the F_1 shoot tip transcriptome (845 P1-dominant genes) and F_2 internode transcriptome (711 P1-dominant genes). The total number of maternally-dominant genes in common among the F_1 and F_2 shoot tip (99) or among the F_1 and F_2 internode (92) transcriptomes was relatively greater than paternally dominant genes shared among the F_1 and F_2 shoot tip (27) or internode (13) transcriptomes.

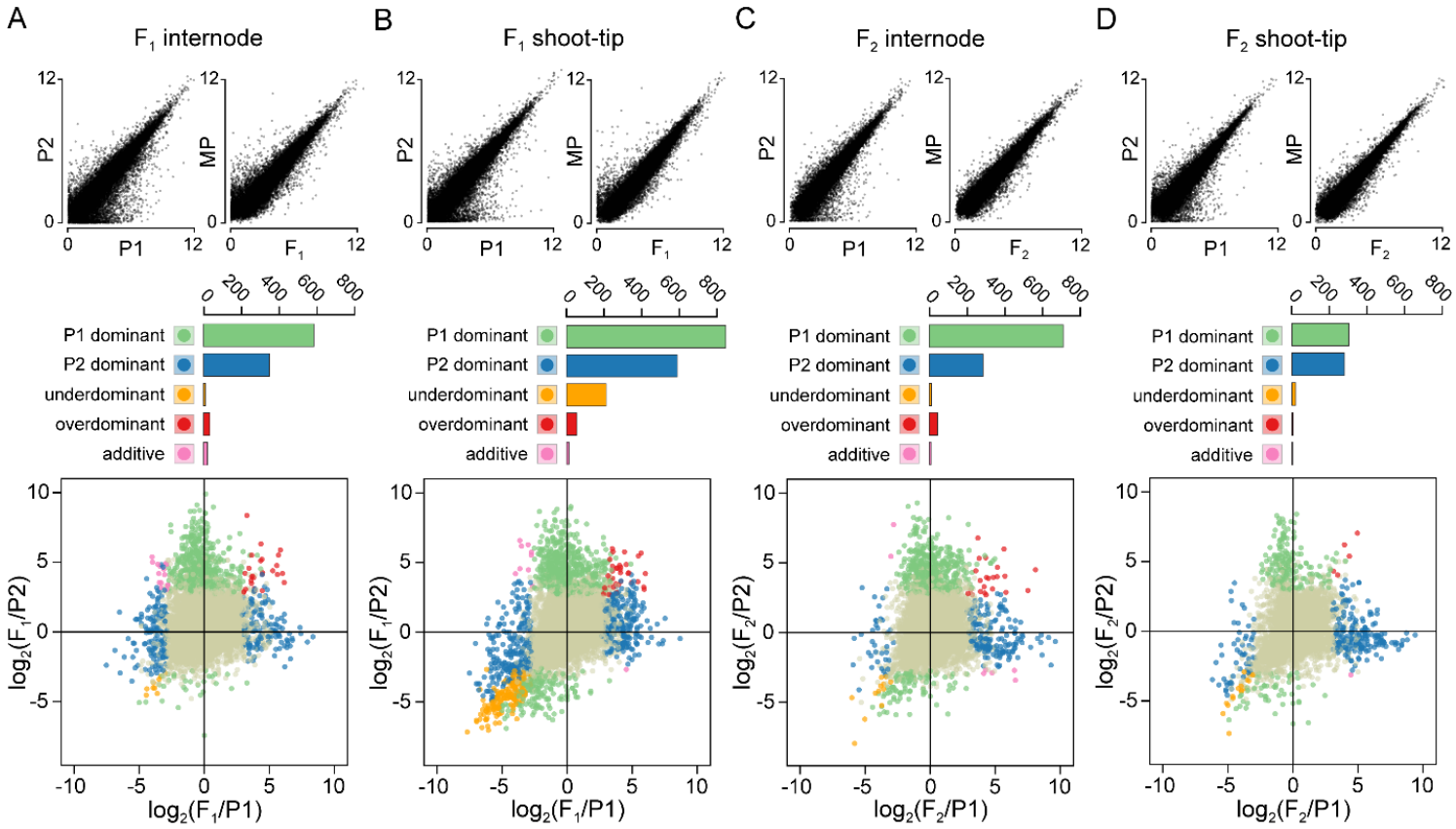

Figure 5.2 Inheritance of gene expression in intraspecific F_1 and F_2 *S. purpurea* families. Pairwise comparison of global gene expression between parents (P1 vs P2) as well as the respective midparent to family progeny (MP vs F_1 or F_2) for each tissue transcriptome. RNA-Seq count data were normalized using $\log_2$ counts-per million mapped reads (cpm) in each library with a prior count of 1. Before normalization, rows with low expression (cpm≤1.0) over 50% of the samples were removed from the analysis. Inheritance of gene expression is summarized in barplots for **(A)** F_1 internode, **(B)** F_1 shoot tip, **(C)** F_2 internode, and **(D)** F_2 shoot tip tissues, and color-coded with respect to classes. Conserved inheritance class (beige) are not shown (reported in Table 4.4). Scatterplots below barplots for each tissue type within a family depict the ratio of $\log_2$ normalized read counts of family progeny to the respective female (x-axis) and male parent (y-axis). Single points within each scatterplot represents unique genes colored according to inheritance classifications (same as boxplots).

Table 5.4 Summary of gene expression inheritance and regulatory divergence classifications of *Salix purpurea* F_1 and F_2 families.

Class	F_1 Family 10X-082				F_2 Family 10X-317			
	Internode		Shoot tip		Internode		Shoot tip	
Expression Inheritance								
P1 dominant	**586**	**2.14%**	**845**	**3.10%**	**711**	**2.63%**	**303**	**1.13%**
P2 dominant	**350**	**1.28%**	**587**	**2.15%**	286	1.06%	279	1.04%
overdominant	34	0.12%	53	0.19%	**43**	**0.16%**	**6**	**0.02%**
underdominant	**10**	**0.04%**	**209**	**0.77%**	12	0.04%	20	0.07%
additive	20	0.07%	13	0.05%	**10**	**0.04%**	**1**	**0.01%**
conserved	**26,441**	**96.4%**	**25,545**	**93.7%**	25,953	96.1%	26,260	97.7%
Total	27,441		27,252		27,015		26,869	
Regulatory Divergence								
cis only	**145**	**2.80%**	**231**	**5.90%**	29	4.40%	41	6.30%
trans only	**85**	**1.60%**	**187**	**4.80%**	28	4.20%	15	2.30%
cis + *trans*	5	0.10%	13	0.30%	2	0.30%	1	0.20%
cis × *trans*	**19**	**0.40%**	**140**	**3.60%**	10	1.50%	6	0.90%
compensatory	**38**	**0.70%**	**512**	**13.0%**	9	1.40%	18	2.80%
ambiguous	833	16.1%	811	20.7%	129	19.4%	100	15.4%
conserved	**4,063**	**78.3%**	**2,033**	**51.8%**	458	68.9%	467	72.1%
Total	5,188		3,927		665		648	

The total number and percentage of genes among those classified within the F_1 and F_2 shoot tip and internode transcriptome are partitioned by their inheritance and regulatory divergence classes (False Discovery Rate=0.005). Numbers in boldface indicate significant (P<0.01) deviations from a 1:1 ratio according to a χ^2 test.

In addition to the extensive expression-level dominance present within these families, >10-fold the number of transgressive genes in the F_1 shoot tip transcriptome were classified as underdominant (209), compared to the F_2 shoot tip transcriptome (20). Likewise, there were fewer genes with overdominant expression in the F_2 shoot tip transcriptome (6) than in the F_1 shoot tip transcriptome (53).

There was a relatively large subset of genes differentially expressed between parents and progeny of the F_1 family (Table 5.5). Many of these genes were classified as transcriptional regulators and hypothetical proteins with domains of unknown function (DUF). Gene ontology (GO) term analysis for P1-dominant genes in internode tissues showed significant enrichment for response to stimulus, catalytic activity and binding, as well as the cellular components intracellular membrane-bound organelle and intracellular part. Significant GO enrichment for P1-dominant genes in the shoot tip was similar to those in the internode for biological processes and cellular components, but included unique molecular functions of catalytic and UDP-glycosyltransferase activity. Enrichment of P2-dominant genes in internode tissues included response to stimulus, transcription regulator activity, and cell wall components, whereas post-embryonic development, programmed cell death, binding, and catalytic activity were over-represented in shoot tip tissue.

Table 5.5 Midparent differentially-expressed genes. Rows within the table are ordered by the -$\log_{10}$(*p*-value) significance of each respective gene. *Salix purpurea* v1 homologs were reported as the best BLAST hit (E ≤ 0.01) to the *Populus trichocarpa* v3 or *Arabidopsis* v10 (TAIR) proteome.

***Salix* gene**	**Location**	**Function**	**Homolog**	**$\log_2$FC**	**$\log_{10}$(*p*)**
Up in Parents					
SapurV1A.3280s0010	scaffold3280	RRM/RBD/RNP motifs family protein	Potri.008G030800	4.7	17.1
SapurV1A.0902s0140	chr01	glycine-rich protein, putative	AT3G17050	4.9	16.2
SapurV1A.0015s0510	chr08	flavin-binding kelch repeat F-box	Potri.008G135200	5.1	13.6
SapurV1A.0050s0340	chr10	plasmodesmata-located 1, putative	Potri.008G104900	4.5	13.2
SapurV1A.0021s0670	chr08	hypothetical protein	AT4G14301	4.4	13
SapurV1A.0741s0010	chr01	glycine/proline-rich protein	Potri.001G141700	4.3	12.8
SapurV1A.1539s0030	scaffold1539	PRR5, pseudo-response regulator 5	Potri.012G005900	3.6	11.5
SapurV1A.0213s0350	chr10	hypothetical protein	Potri.010G230400	5.1	9.7
SapurV1A.1507s0010	scaffold1507	transmembrane protein, putative	Potri.001G141700	3.6	9.5
SapurV1A.0857s0090	chr01	hypothetical protein	Potri.001G021300	3.5	9.4
SapurV1A.0014s1200	chr02	phragmoplast kinesin 2-like	AT3G19050	4.2	8.8
SapurV1A.1107s0100	chr13	RELA/SPOT homolog 2	AT3G14050	3.5	8.3
SapurV1A.0741s0020	chr01	transmembrane protein, putative	Potri.001G141800	3.3	8.2
SapurV1A.0402s0140	chr02	Ole e 1, pollen protein	AT2G43150	3.0	8.0
SapurV1A.1734s0030	chr02	trichome biferingence-like 32	AT3G11030	3.8	7.7
SapurV1A.0808s0130	chr07	hypothetical protein		3.0	6.6
SapurV1A.1897s0030	chr02	hypothetical protein		3.4	6.5
SapurV1A.1522s0030	scaffold1522	terpene synthase, putative	Potri.019G023100	3.4	6.5
SapurV1A.0132s0300	chr01	patatin-like phospholipase		3.5	6.2
SapurV1A.3446s0010	scaffold3446	short-chain dehydrogenase reductase	Potri.015G104800	3.8	6.1
SapurV1A.0850s0060	scaffold0850	JMJ30, jumonji-C domain protein 30	Potri.001G016200	2.9	6.1
SapurV1A.1733s0040	scaffold1733	protodermal factor 1.1, putative	Potri.002G060800	2.9	6.1
SapurV1A.0802s0230	chr01	cytochrome P450 family protein	Potri.001G054800	3.3	5.9
SapurV1A.0306s0280	chr06	hypothetical protein		3.1	5.9
SapurV1A.0500s0140	chr19	transcription factor TCP4	Potri.019G091300	2.9	5.9

Table 5.5
(continued)

***Salix* gene**	**Location**	**Function**	**Homolog**	**$\log_2$FC**	**$\log_{10}(p)$**
SapurV1A.0313s0200	scaffold0313	GDSL-like lipase/acylhydrolase	Potri.018G089200	2.2	5.4
SapurV1A.3280s0020	scaffold3280	hypothetical protein	Potri.008G030800	7.0	5.3
SapurV1A.5565s0010	scaffold5565	kinesin motor-like protein	Potri.016G060400	7.0	5.3
SapurV1A.0937s0110	chr02	transmembrane protein, putative	Potri.002G252400	3.4	5.3
SapurV1A.1305s0040	scaffold1305	KPI-A2, kunitz protease inhibitor		3.1	5.3
SapurV1A.0066s0170	chr13	hypothetical protein	Potri.002G070100	2.6	5.2
SapurV1A.2115s0020	scaffold2115	hypothetical protein		2.2	5.2
SapurV1A.2259s0020	chr18	transmembrane protein, putative	Potri.018G108700	3.5	5.1
SapurV1A.3047s0030	chr13	hypothetical protein		2.8	5.1
SapurV1A.1313s0010	chr17	hypothetical protein	Potri.017G047000	3.3	4.9
SapurV1A.0056s0220	chr03	inhibitor/seed storage/LTP protein	Potri.003G172400	2.6	4.8
SapurV1A.0611s0190	chr14	growth-regulating factor 1, putative	Potri.014G012800	2.4	4.8
SapurV1A.4399s0010	chr01	lipoxygenase	Potri.001G015400	2.9	4.7
SapurV1A.0685s0220	chr19	KIN11, SNF1-related protein kinase	AT3G29160	2.6	4.7
SapurV1A.4611s0010	scaffold4611	transmembrane protein, putative	Potri.015G060000	6.6	4.6
SapurV1A.1353s0050	chr13	cytochrome P450 family protein	Potri.013G157200	3.6	4.5
SapurV1A.0946s0070	chr13	GDSL esterase/lipase	Potri.019G024400	2.7	4.5
SapurV1A.0258s0090	chr16	YABBY2, axial regulator		2.5	4.5
SapurV1A.1459s0030	chr03	hypothetical protein		2.1	4.5
SapurV1A.1439s0010	chr04	RBM42, RNA-binding protein 42	Potri.004G155300	2.2	4.4
SapurV1A.6991s0010	scaffold6991	ATP-dependent Clp protease subunit	Potri.013G017200	3.1	4.3
SapurV1A.0429s0110	chr04	transmembrane protein, putative	Potri.004G009900	2.8	4.3
SapurV1A.0083s0520	chr16	TCP family transcription factor	Potri.001G375800	2.2	4.3
SapurV1A.0446s0100	chr06	hypothetical protein	Potri.006G240400	2.1	4.3
SapurV1A.1301s0110	scaffold1301	hypothetical protein		2.9	4.2
SapurV1A.0722s0230	chr04	homeobox leucine zipper protein	Potri.004G020400	2.5	4.2
SapurV1A.0191s0130	chr17	tryptophan synthase beta chain	Potri.017G109600	2.3	4.2
SapurV1A.0280s0060	chr01	hypothetical protein		6.4	4.1

Table 5.5
(continued)

***Salix* gene**	**Location**	**Function**	**Homolog**	**$\log_2$FC**	**$\log_{10}(p)$**
SapurV1A.1274s0030	scaffold1274	MYB transcription factor, putative		2.7	4.1
SapurV1A.0268s0230	chr15	transmembrane protein, putative		2.5	4.1
SapurV1A.2659s0010	scaffold2659	dehydrin protein		2.5	4.1
SapurV1A.1549s0040	chr04	cytochrome P450 family protein	Potri.004G019000	2.3	4.1
SapurV1A.0032s0280	chr11	Ole e 1, pollen protein	Potri.011G053600	2.3	4.1
SapurV1A.0277s0090	chr13	ZF-HD homeobox protein	Potri.013G108900	2.3	4.1
SapurV1A.1443s0020	chr19	GDSL esterase/lipase	Potri.013G051000	2.3	4.1
SapurV1A.0272s0100	chr06	carotenoid cleavage dioxygenase	Potri.006G238500	3.2	4.0
SapurV1A.0316s0210	chr14	transmembrane protein, putative	Potri.002G227700	2.9	4.0
SapurV1A.2732s0020	scaffold2732	hypothetical protein	Potri.019G105300	2.8	4.0
SapurV1A.0034s0500	chr18	YABBY2, axial regulator	Potri.018G129800	2.5	4.0
SapurV1A.0608s0150	chr07	growth-regulating factor, putative	Potri.007G007100	2.0	4.0
SapurV1A.0014s0820	chr02	hypothetical protein		2.6	3.9
SapurV1A.2014s0020	scaffold2014	dehydrin protein		2.3	3.9
SapurV1A.1480s0080	chr04	UDP-glycosyltransferase	Potri.004G070000	1.8	3.9
Up in F_1					
SapurV1A.0085s0560	chr05	CNGC16, cyclic nt-gated channel 16	AT3G48010	4.8	11.5
SapurV1A.1864s0010	chr03	K-box DENN (AEX-3) domain	AT5G35560	4.0	6.6
SapurV1A.0289s0100	chr02	GRAS family transcription factor	AT3G50650	5.5	5.4
SapurV1A.0173s0350	chr03	MADS-box transcription factor	Potri.003G169800	3.6	5.3
SapurV1A.0464s0250	chr06	ERD15, polyadenylate-binding protein	Potri.006G044600	3.0	5.2
SapurV1A.0455s0050	chr07	ROOT HAIR defective-like protein		3.4	5.1
SapurV1A.0833s0010	chr16	MADS-box transcription factor		3.1	4.9
SapurV1A.1182s0050	chr05	MADS-domain transcription factor		6.2	4.3

5.4.3 ASE Analysis

In order to discern the overall proportion of *cis*- and *trans*-regulation of gene expression in *S. purpurea*, ASE tests were conducted using RNA-Seq expression data which was based on biallelic sites called from DNA-Seq and RNA-Seq of the parents. For both families, the regulation of gene expression was primarily conserved, regardless of the tissue type assayed (Figure 5.3). However, for those genes showing non-conserved regulatory classes, moderate proportions of pure *trans*-regulated gene expression were identified in the F_1 shoot tip (187, 11%) and internode (85, 14.5%) transcriptome, but substantially less in the F_2 shoot tip (15, 2.5%) and internode (28, 2.6%) transcriptome (Table 5.6). The proportion of ASE in the F_1 shoot tip transcriptome was nearly twice that of the F_1 internode transcriptome, yet there was no difference among F_2 tissues. On average, the F_1 had a greater number of SNPs per gene than the F_2 (Figure 5.3; Table 5.6). As the number of SNPs per gene increased, the $\log_2$ (P1/P2) expression ratio decreased. The only major discrepancy in the average number of SNPs per gene was between F_1 tissues for those genes showing *cis* + *trans* regulatory interactions.

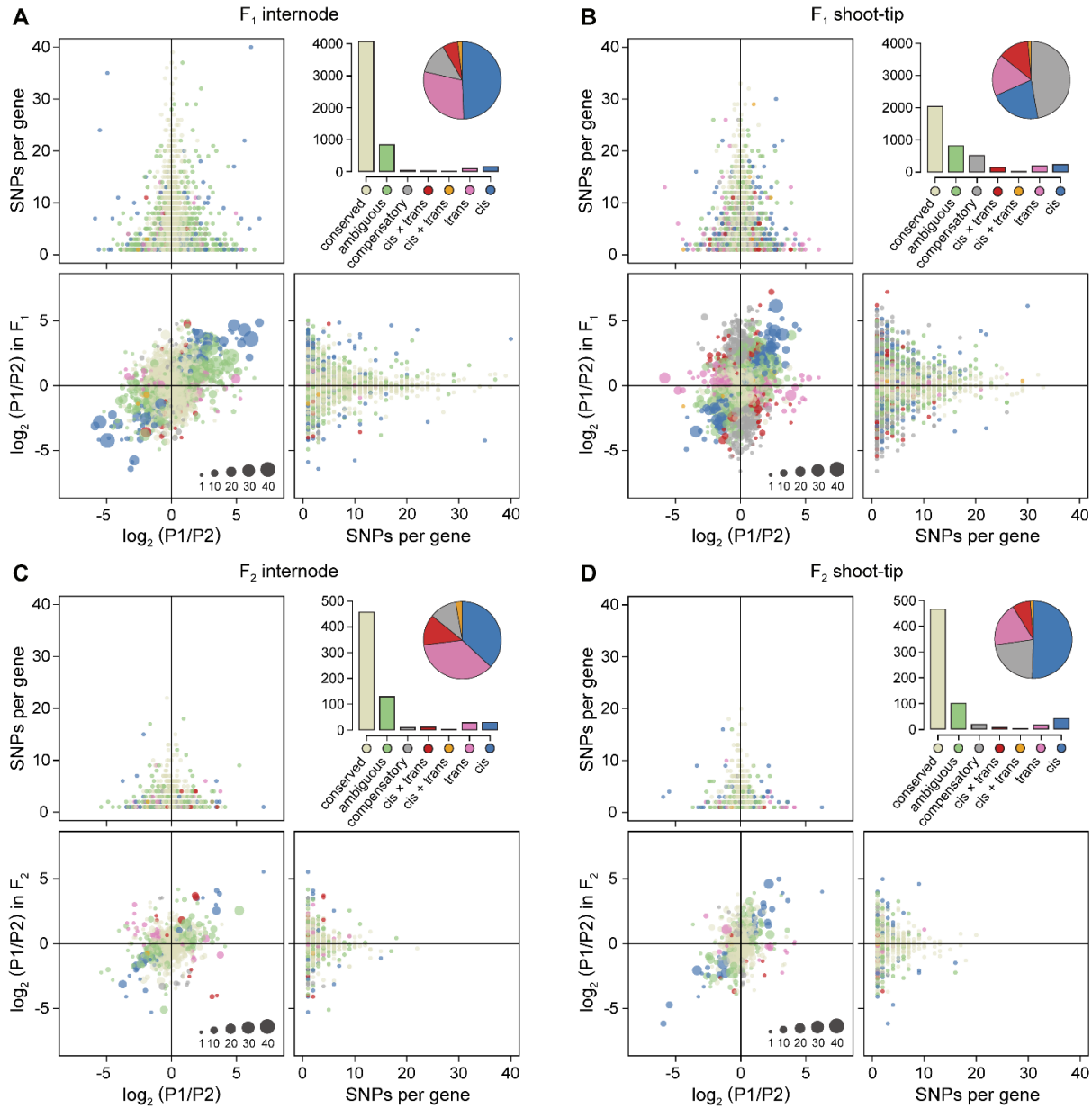

Figure 5.3 Allele-specific expression in intraspecific F_1 and F_2 *S. purpurea* families. Regulatory divergence classifications are summarized in barplots and proportions in pie charts for the **(A)** F_1 internode, **(B)** F_1 shoot tip, **(C)** F_2 internode, and **(D)** F_2 shoot tip tissues. Scatterplots in lower left panel for each tissue transcriptome within a family depict regulatory divergence as a ratio of allele-specific expression of the parents to that of F_1 and F_2 family progeny. For panels **A-D**, the number of SNPs is plotted against ASE ratios. Single points within each scatterplot represents unique genes colored according to respective regulatory classifications and scaled using the log number of SNPs per gene. Regulatory divergence assignments were based on binomial exact tests performed between the female parent (P1) and the male parent (P2) and Fisher's Exact test of the female and male parent alleles in the hybrid at an FDR global significance threshold of 0.005.

Table 5.6 Average number of SNPs per gene by regulatory class.

	F_1 Family 82		**F_2 Family 317**	
	Internode	**Shoot tip**	**Internode**	**Shoot tip**
cis only	4.4	4.1	2.4	3.3
trans only	2.8	3.0	2.0	2.2
cis + *trans*	2.2	5.8	1.5	1.0
cis × *trans*	2.7	2.0	2.0	1.2
compensatory	1.7	2.7	2.8	1.8
ambiguous	3.5	3.8	1.6	2.5
conserved	3.6	4.4	2.9	3.1
average	3.0	3.7	2.2	2.1

A significantly greater number genes showing *cis*- and *trans*-regulation in the F_1 coincided with upregulation of the maternal P1 allele, irrespective of the magnitude, whereas the paternal P2 allele was higher expressed for genes showing *cis* × *trans* or compensatory regulation. Compensatory regulation accounted for 13% (512) of gene expression in the F_1 shoot tip transcriptome, but only 0.7% (38) in the internode transcriptome. Without considering conserved or ambiguous classes in the F_1 shoot tip transcriptome, 30% exhibited compensatory patterns and 8.3% with *cis* × *trans* regulation, compared to 13.5% and 11% of genes with pure *cis*- and *trans*-regulation, respectively.

Further analysis of the F_1 revealed significantly greater levels of nonadditive and regulatory divergent expression in the shoot tip transcriptome compared to the internode transcriptome. Among the unpooled libraries of F_1 progeny individuals (Figure 5.4), the degree of midparent differential expression corresponded linearly to the variation in regulatory divergent expression, whereby increased levels of compensatory expression coincided with increased levels of *cis* × *trans* regulation (Figure 5.5). However, compensatory regulation was negligible in the F_2 family and did not show significant differences by tissue type.

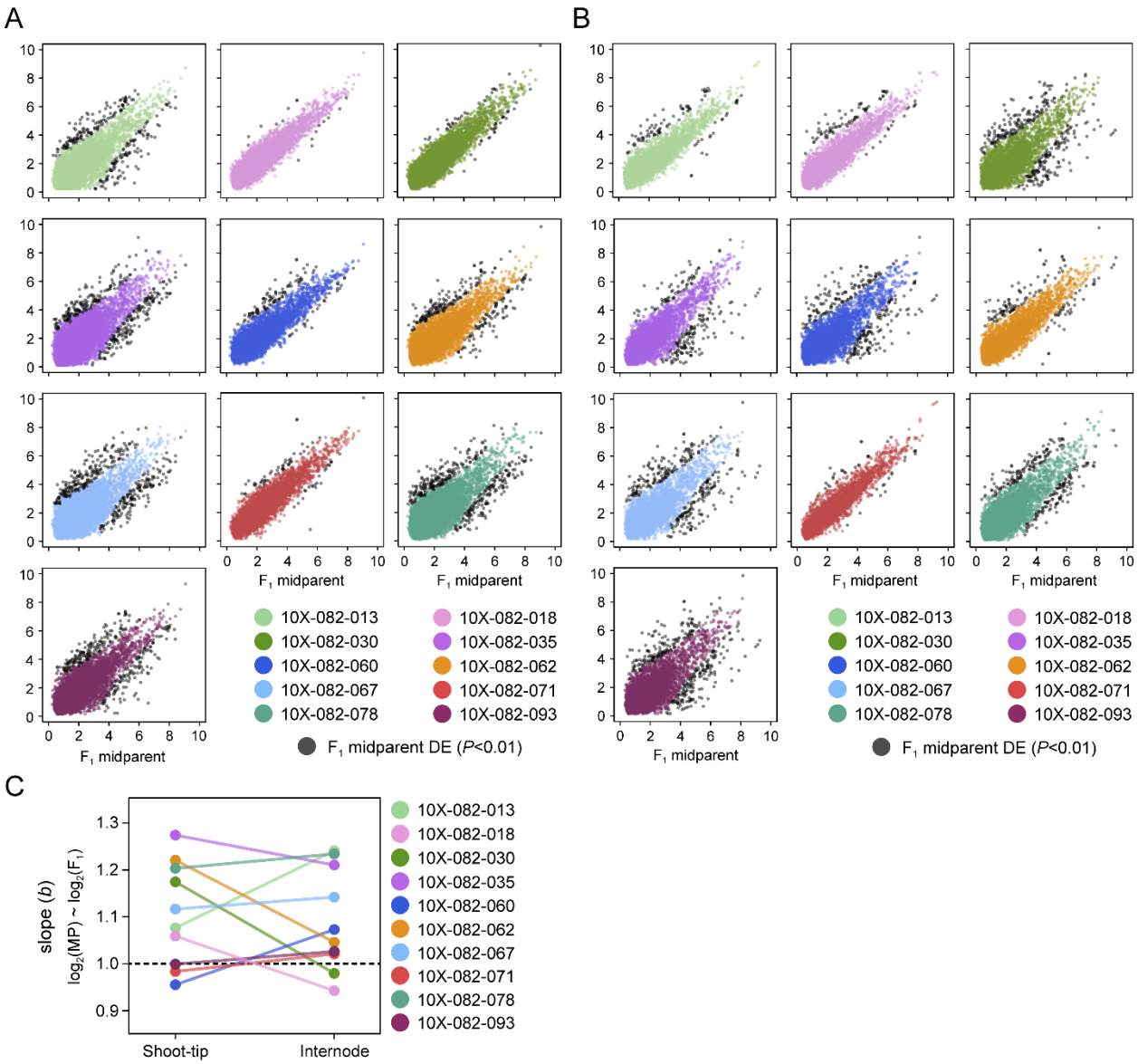

Figure 5.4 Midparent differential expression of F_1 individuals. Genes with significant midparent differential expression (grey points) are highlighted in each panel for **(A)** shoot tip and **(B)** internode transcriptome of F_1 individuals. The line chart in **(C)** depicts the slope of the regression of log_2-normalized gene expression of the F_1 midparent and 10 F_1 progeny individuals as individual points for shoot tip and internode tissues. For each F_1 individual, a line connects the slope (points) for each tissue-type. The dotted horizontal black line represents a slope of 1. Color coding for each F_1 individual is shown to the right.

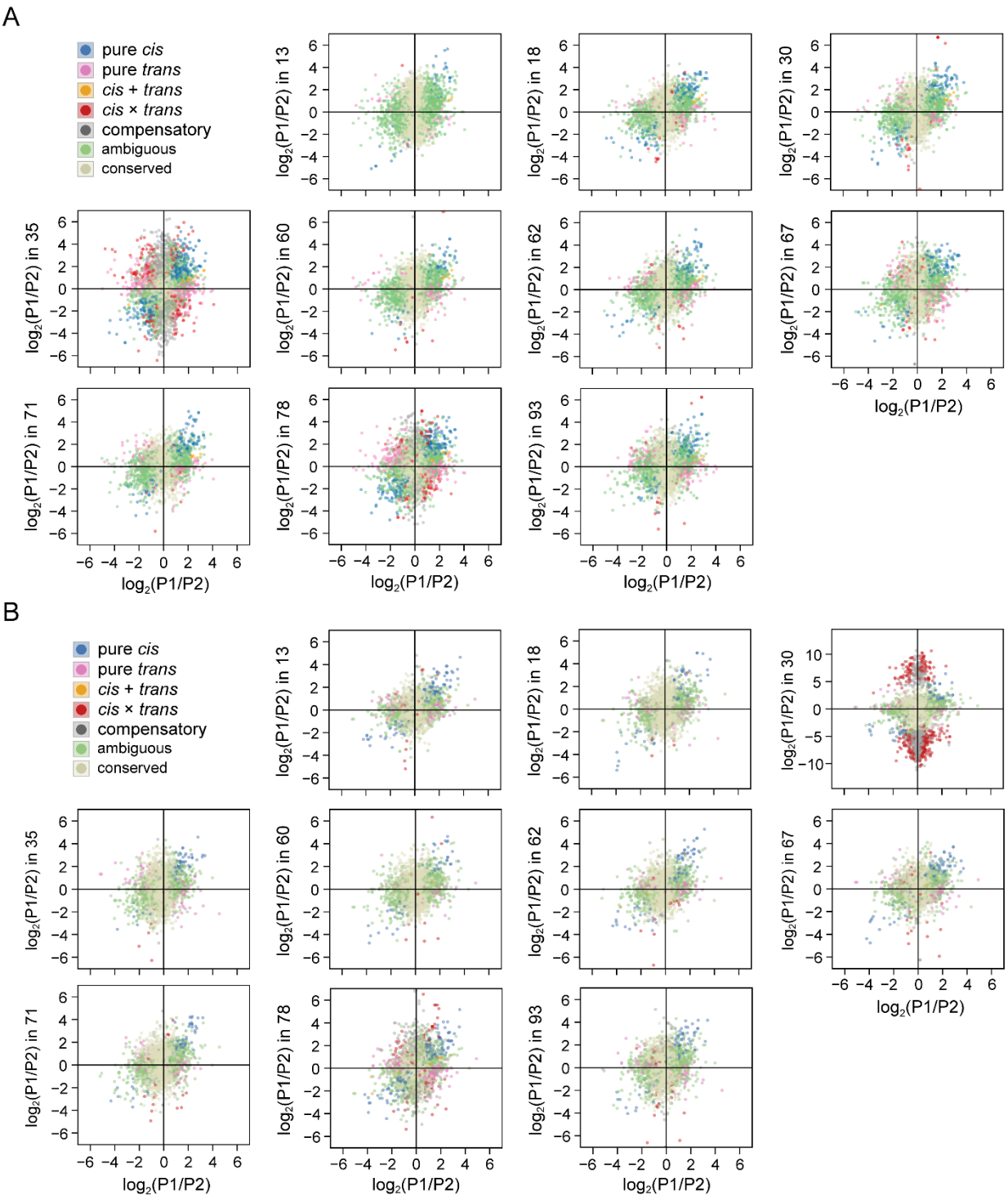

Figure 5.5 Allele-specific expression and regulatory divergence of F_1 individuals. Regulatory divergent expression of **(A)** shoot tip and **(B)** internode tissues for 10 F_1 progeny individuals. Color coding for regulatory classifications are shown above each panel.

5.4.4 Tissue-Biased Gene Expression

Individuals within the F_1 family were independently tested to investigate the average effect of tissue type. After library normalization and filtering for low expression, we identified a total of 262 genes as differentially expressed between the F_1 shoot tip and internode transcriptome. A subset of 46 genes were compiled (Table 5.7) that were highly expressed in internode tissues with $\log_2$ fold-differences ranging from 3.2 (SapurV1A.0130s0080) to 6.0 (SapurV1A.0216s0270). Genes encoding for fasiclin-like arabinogalactan (FLA) proteins were most represented. A total of 20 FLAs from the 57 FLA gene family members were identified as differentially expressed (Figure 5.6), all of which are annotated *FLA11* or *FLA12* with a single *FLA2* representative (SapurV1A.0054s0430). GO term analyses of internode predominant genes showed significant enrichment for cell wall and biosynthetic processes.

Table 5.7 Differentially-expressed genes among shoot-tip and internode tissues. Rows within the table are ordered by the -$\log_{10}$(*p*-value) significance of each respective gene. *Salix purpurea* v1 homologs were reported as the best BLAST hit ($E \leq 0.01$) to the *Populus trichocarpa* v3 or *Arabidopsis* v10 (TAIR) proteome.

***Salix* gene**	**Location**	**Function**	**Homolog**	**$\log_2$FC**	**$\log_{10}(p)$**
Up in Internode					
SapurV1A.0216s0270	chr04	FLA12, fasciclin-like AGP protein	Potri.004G210600	6.0	13.4
SapurV1A.0216s0260	chr04	FLA12, fasciclin-like AGP protein	Potri.004G210600	5.4	10.8
SapurV1A.1427s0060	chr19	FLA12, fasciclin-like AGP protein	Potri.019G123100	5.6	10.5
SapurV1A.5529s0010	chr09	FLA12, fasciclin-like AGP protein	Potri.009G012200	5.0	10.5
SapurV1A.0174s0090	chr12	FLA12, fasciclin-like AGP protein	Potri.012G015000	5.0	9.4
SapurV1A.0524s0120	chr15	FLA12, fasciclin-like AGP protein	Potri.015G013300	5.4	9.0
SapurV1A.0216s0050	chr04	epithiospecifier pseudogene, putative	Potri.004G212900	5.5	8.6
SapurV1A.1289s0040	chr06	FLA12, fasciclin-like AGP protein	Potri.009G012200	4.5	8.4
SapurV1A.2928s0040	scaffold2928	FLA12, fasciclin-like AGP protein	Potri.013G151300	5.1	8.3
SapurV1A.0604s0140	chr13	FLA12, fasciclin-like AGP protein	Potri.013G014200	4.5	8.3
SapurV1A.0433s0060	chr15	FLA12, fasciclin-like AGP protein	Potri.019G121300	4.3	8.3
SapurV1A.0069s0100	chr08	hypothetical protein	n/a	4.7	8.2
SapurV1A.0300s0070	chr17	MIF2, mini ZF-HD protein	Potri.004G126600	4.5	8.2
SapurV1A.0433s0050	chr15	FLA12, fasciclin-like AGP protein	Potri.019G123100	4.3	8.2
SapurV1A.0113s0470	chr08	RD22, BURP domain protein	Potri.T012700	3.6	8.0
SapurV1A.0318s0140	chr07	CYP96A, cytochrome P450 protein	Potri.007G080300	4.7	7.7
SapurV1A.0007s1240	chr10	PHO1, phosphate transporter Pho1	Potri.008G169400	5.0	7.4
SapurV1A.0041s0490	chr16	hypothetical protein	Potri.001G245400	4.3	7.0
SapurV1A.0843s0100	chr17	TPX2, targeting protein for Xklp2	Potri.T059900	3.8	7.0
SapurV1A.0235s0080	chr14	rhamnogalacturonan endolyase	Potri.014G004500	4.8	6.7
SapurV1A.0171s0440	chr07	MYB69, Myb transcription factor	Potri.007G106100	4.5	6.6
SapurV1A.0262s0360	chr10	protease inhibitor/seed storage protein	Potri.010G196300	3.6	6.5
SapurV1A.0035s0160	chr01	SWEET4, sugar transporter	Potri.003G143100	4.8	6.4
SapurV1A.1443s0050	chr19	GDSL esterase/lipase	Potri.019G024800	4.4	6.4
SapurV1A.0311s0290	chr05	AAP2, amino acid permease	Potri.005G068900	4.2	6.4

Table 5.7
(continued)

***Salix* gene**	**Location**	**Function**	**Homolog**	**log_2FC**	**$log_{10}(p)$**
SapurV1A.0829s0140	scaffold0829	ACC oxidase	Potri.002G078600	4.1	6.4
SapurV1A.2928s0010	scaffold2928	FLA12, fasciclin-like AGP protein	Potri.013G014200	3.5	6.4
SapurV1A.1100s0020	chr06	TUBC2, tubby C 2 protein	Potri.006G067000	5.0	6.3
SapurV1A.1361s0090	chr05	cyclin-dependent kinase, putative	Potri.005G033600	4.7	6.3
SapurV1A.0001s1660	chr06	LAC2, laccase 2	Potri.006G087100	3.5	6.3
SapurV1A.0174s0080	chr12	FLA12, fasciclin-like AGP protein	Potri.012G015000	4.2	6.2
SapurV1A.0323s0110	scaffold323	IRX6, COBRA-like protein	Potri.004G117200	4.4	6.1
SapurV1A.0130s0080	chr05	PER64, peroxidase 64	Potri.005G108900	3.2	6.1
SapurV1A.0587s0110	chr18	2OG-Fe(II) oxidoreductase	Potri.018G121800	4.6	6.0
SapurV1A.0428s0120	chr5	peroxidase	Potri.005G135300	4.5	6.0
SapurV1A.0248s0100	chr14	UNE11, pectin methylesterase inhibitor	Potri.014G067500	4.7	5.9
SapurV1A.0107s0230	chr03	transmembrane protein, putative	Potri.003G083000	4.5	5.9
SapurV1A.2107s0040	scaffold2107	zinc finger protein	Potri.013G060500	4.5	5.8
SapurV1A.0079s0510	chr08	hypothetical protein	n/a	4.4	5.8
SapurV1A.4071s0010	chr13	FLA12, fasciclin-like AGP protein	Potri.013G151300	4.4	5.8
Up in Shoot-tip					
SapurV1A.1011s0090	chr19	PLP5, patatin-like phospholipase	AT4G37060	7.4	11.5
SapurV1A.2520s0030	chr16	PRB1, pathogenesis-related protein	Potri.T131500	6.6	9.7
SapurV1A.1341s0020	chr15	LCR69, gamma-thionin, putative	Potri.T011200	3.6	9.4
SapurV1A.1263s0010	scaffold1263	ST2A, flavonol 4'-sulfotransferase	Potri.003G189100	5.0	8.5
SapurV1A.1011s0040	chr19	PLP5, patatin-like phospholipase	AT4G37060	5.9	6.3
SapurV1A.0751s0210	chr15	CHIA, acidic chitinase	Potri.015G024000	4.0	6.3
SapurV1A.0946s0020	chr13	GDSL-lipase	Potri.018G089100	4.8	5.4
SapurV1A.0436s0110	scaffold0436	hypothetical protein	AT2G26560	5.0	5.2
SapurV1A.1263s0060	scaffold1263	ST2A, flavonol 4'-sulfotransferase	Potri.003G189100	4.4	5.2
SapurV1A.1781s0040	scaffold1781	SWEET4, sugar transporter	AT5G24090	3.4	4.9
SapurV1A.0500s0140	chr19	TCP4, transcription factor TCP4	Potri.019G091300	3.2	4.8

Table 5.7
(continued)

***Salix* gene**	**Location**	**Function**	**Homolog**	**$\log_2$FC**	**$\log_{10}(p)$**
SapurV1A.0258s0090	chr16	YABBY2, transcription factor YAB2	AT2G26580	3.1	4.8
SapurV1A.1188s0120	chr04	hypothetical protein	n/a	3.1	4.4
SapurV1A.1954s0020	chr13	SWEET4, sugar transporter	Potri.013G013800	4.1	3.9
SapurV1A.0164s0090	chr08	asymmetric leaves protein, putative	Potri.008G079800	3.1	3.9
SapurV1A.0279s0070	chr16	PRB1, pathogenesis-related protein	Potri.T131500	3.3	3.7
SapurV1A.0522s0010	chr19	Wax ester synthase	AT5G53390	3.2	3.7
SapurV1A.1762s0060	scaffold1762	mTERF protein	Potri.004G012400	3.6	3.5
SapurV1A.1668s0020	chr15	ribonuclease P/MRP POP5	Potri.015G001200	3.3	3.5

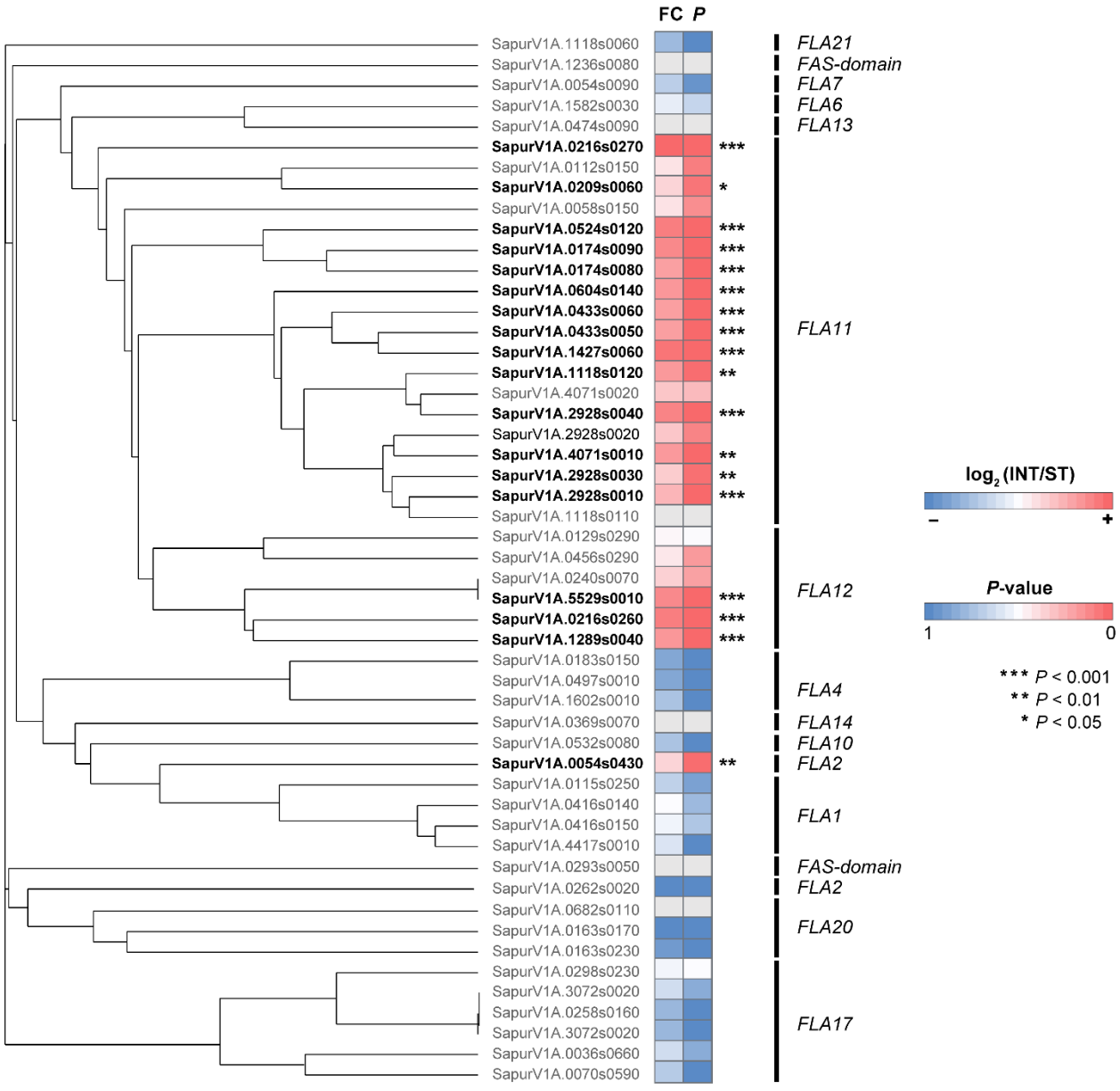

Figure. 5.6 Neighbor-Joining Tree and differential expression of the fasciclin gene family in *Salix*. The column labeled FC represents the $\log_2$ fold-change in cpm-normalized expression of internode and shoot tip tissues. Differential expression is denoted by asterisks ***, **, and * beside the *P* column header, significant at an adjusted *p*-value < 0.1%, 1%, and 5%, respectively. A total of 20 FLAs from 57 FLA gene family members were identified as differentially expressed. All differentially-expressed genes but SapurV1A.0054s0430, a *FLA2* representative, are annotated as either *FLA11* or *FLA12* family members.

5.4.5 Sex-Biased Gene Expression

In order to test for differential expression between male and female *S. purpurea*, we utilized the shoot tip transcriptome of three male and three female F_1 individuals as well as their parents, such that each sex was represented by four related individuals. We identified a total of 315 genes in the F_1 shoot tip transcriptome as having significant sex-biased expression (Table 5.8). In stark contrast, there were no genes in the F_1 internode transcriptome that showed significant sex-biased expression. Of the 315 sex-biased genes, 62 map to *S. purpurea* chr15. In addition, 77 genes with best BLAST hits to *P. trichocarpa* v3 chr15 orthologues accounted for ~24% of the sex-biased genes identified (Figure 5.7B). From this list, 231 genes were more highly expressed in females and 84 were more highly expressed in males, indicating that a significantly higher proportion of the shoot tip transcriptome is female-biased than male-biased (Figure 5.7C). Nearly all female-biased genes reported in this study localize to *Salix* chr15 or unplaced scaffolds syntenic to *Populus* chr15. By contrast, male-biased genes were not predominantly on *S. purpurea* chr15, many mapped to *S. purpurea* chr19, especially for genes with higher expression in males. Over half of the sex-biased genes on *S. purpurea* chr19 which were highly expressed in males encode proteins related to signaling and response (e.g. ankyrin repeat, patatin phospholipase, and nucleotide-binding site leucine-rich repeat proteins), but do not appear to localize to any particular region of the chromosome.

Table 5.8 Sex-biased gene information.

Salix gene	Chr	Functional annotation	Homolog	log_2	$-log_{10}(p)$
Up in Females					
SapurV1A.0582s0010	582	NIPA, interacting partner of ALK	Potri.015G052400	6.7	42.6
SapurV1A.2504s0020	2504	AGL98, agamous 98-like	AT5G39810	7.7	39.9
SapurV1A.4040s0010	4040	Di-glucose binding, kinesin	Potri.001G436200	3.3	27.3
SapurV1A.0301s0070	15	REM1, reproductive meristem 1	Potri.009G103300	8.8	25.3
SapurV1A.2504s0010	2504	GPI-anchored protein	Potri.015G040900	6.0	24.6
SapurV1A.0301s0160	15	DR1/NF-YB, TBP-associated	Potri.015G052800	4.7	21.0
SapurV1A.0301s0170	15	BRCA, fragile-X-F-associated	Potri.015G050300	6.8	20.8
SapurV1A.0475s0170	15	peptidase M50B-like protein	Potri.015G045900	9.8	18.2
SapurV1A.1892s0010	15	MOS4/SPF27, modifier of SNC1	Potri.015G041800	5.1	17.0
SapurV1A.2524s0010	2524	PHYB, phytochrome protein B	Potri.008G105200	4.1	13.7
SapurV1A.2212s0030	2212	activating signal cointegrator 1,3	Potri.015G056700	5.5	12.5
SapurV1A.4349s0010	4349	SCD1, cytokinesis-defective 1	Potri.015G049500	3.9	11.7
SapurV1A.1210s0090	1210	PME36, pectinesterase inhibitor 36	Potri.015G127700	6.9	11.6
SapurV1A.1386s0030	15	HOT101, heat shock protein 101	Potri.015G056900	3.1	10.3
SapurV1A.0530s0090	15	LRK10, serine/threonine kinase	Potri.015G044800	4.7	9.9
SapurV1A.0178s0110	15	18S pre-ribosomal, gar2-related	Potri.015G048400	5.9	9.2
SapurV1A.1146s0050	15	LP-1, thaumatin protein 1	Potri.015G039200	2.2	9.2
SapurV1A.0301s0080	15	CaS, extracellular Ca^{2+} receptor	Potri.015G052200	3.2	8.8
SapurV1A.0107s0110	3	UBC2/RAD6, Ub conjugating E2, 1	Potri.013G064400	4.0	8.8
SapurV1A.0582s0060	582	GUS2/HPSE1, heparanase 1-like	Potri.015G049100	5.0	8.3
SapurV1A.0530s0130	15	RTNLB9, reticulon-like protein	Potri.015G044300	7.9	7.8
SapurV1A.2212s0020	2212	activating signal cointegrator 1,3	Potri.015G056500	5.0	7.6
SapurV1A.0107s0070	3	IMPA-2, importin alpha 2	Potri.005G020400	4.5	7.5
SapurV1A.0107s0060	3	GATA Znf protein	Potri.005G020500	2.8	7.0
SapurV1A.1538s0020	15	TCP-1/cpn60, delta chaperonin	Potri.015G042600	5.6	6.7
SapurV1A.1002s0030	15	WOX5, wuschel-related homeobox 5	Potri.015G065400	2.9	6.6
SapurV1A.0530s0070	15	delta-ADR, AP-3 complex delta-1	Potri.015G045600	2.5	6.5
SapurV1A.1254s0040	19	AMY-1, associate of c-MYC	Potri.019G014100	7.1	5.9

Table 5.8
(continued)

Salix gene	Chr	Functional annotation	Homolog	$\log_2$	$-\log_{10}(p)$
SapurV1A.2772s0010	15	RPL19e/EMB2386, ribosomal 19e	Potri.015G037100	4.7	5.9
SapurV1A.0582s0100	582	LAP4, less-adhesive pollen 4	Potri.015G048800	5.2	5.6
SapurV1A.0582s0040	582	meiotic endonuclease, putative	Potri.015G049800	6.7	5.5
SapurV1A.2535s0010	2535	suppressor of protein silencing	Potri.018G137400	6.6	5.4
SapurV1A.0178s0160	15	RING/U-box Znf protein	Potri.015G047900	2.6	5.4
SapurV1A.1596s0050	15	DYNLL1, dynein light chain 1-like	Potri.015G067800	4.6	4.8
SapurV1A.0307s0060	19	PIF1, phytochrome interacting 1	Potri.008G203700	7.3	4.6
Up in Males					
SapurV1A.0934s0010	15	RPS3, 40S ribosomal protein S3-1	Potri.015G071700	1.9	6.4
SapurV1A.0830s0010	830	NIPA, interacting partner of ALK	Potri.015G052400	1.2	6.1
SapurV1A.0934s0060	15	DR1/NF-YB, TBP-associated	Potri.015G052800	1.3	5.5
SapurV1A.3555s0010	3555	CLO1-2, caleosin 1, seed gene 1	Potri.010G066600	1.3	5.1
SapurV1A.1765s0050	1765	fertility restorer (Rf)	Potri.015G036400	1.4	4.8
SapurV1A.0530s0040	15	peptidase M50B	Potri.015G045900	1.3	3.9
SapurV1A.1246s0030	15	transmembrane protein	AT3G18215	1.3	3.9
SapurV1A.1510s0020	15	GPI-anchored protein	Potri.015G040900	1.1	3.8
SapurV1A.0391s0140	19	ankyrin repeat, SAM domain 1	Potri.019G106200	6.7	3.8
SapurV1A.0704s0100	15	TB2/DP1, HVA22 family protein	Potri.015G062800	1.8	3.8
SapurV1A.0391s0170	19	ankyrin repeat protein	Potri.019G107700	6.3	3.6
SapurV1A.1421s0010	15	NOF1/Utp25, nucleolar factor 1	Potri.003G010000	5.0	3.5
SapurV1A.1515s0010	15	CAAX amino terminal protease	Potri.019G101100	2.7	3.5

Rows within the table are ordered by the $-\log_{10}(p\text{-value})$ significance of each respective gene. *Salix* v1 homologs are reported as the best BLAST hit (E-value≤0.01) to the *Populus trichocarpa* v3 or Arabidopsis TAIR v10 proteome.

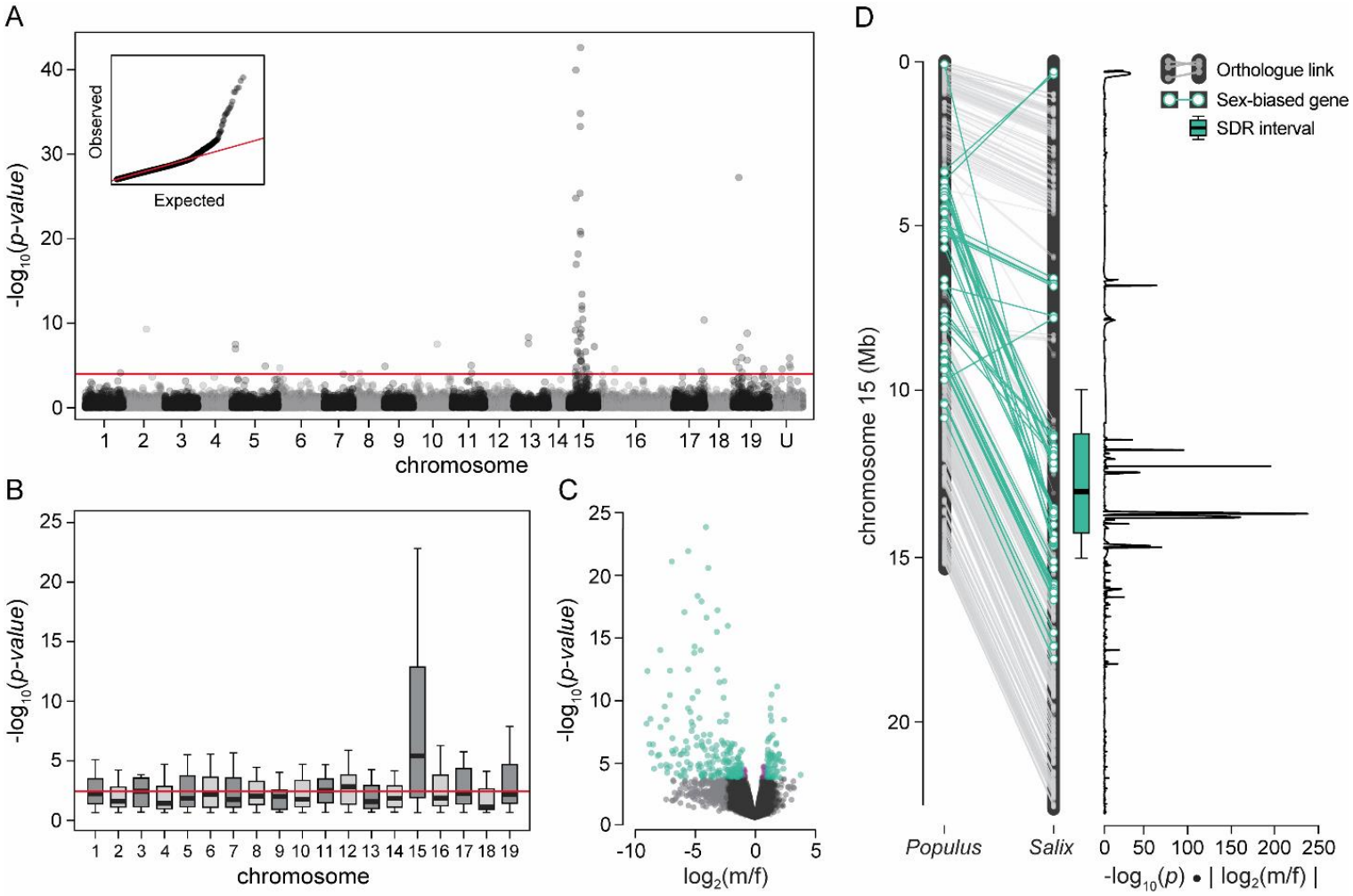

Figure 5.7 Sex-biased expression maps to *Salix purpurea* chr15. Significant sex-biased gene expression is depicted in the Manhattan plot (**A)**, where each point represents the -$\log_{10}$ of the adjusted *p*-value for each gene. Chromosomes are ordered from 1-19 (unplaced scaffold IDs not shown), the horizontal red line represents the global significance threshold, and QQ-plot showing model-fit in the upper left. Within each boxplot demonstrating the chromosomal distribution of differentially expressed genes **(B)**, the solid black line represents the median -$\log_{10}(p)$ for each chromosome with whiskers extending an interquartile range of 1.5, and the median

Figure 5.7
(continued)

genome-wise -$\log_{10}(p)$ is represented by the red horizontal line. A volcano plot **(C)** depicts the magnitude of the $\log_2$(m/f) change in gene expression (x-axis) and the -$\log_{10}(p)$ significance (y-axis) for sex, where negative values depict upregulated genes in females and positive values, upregulation in males. The magnitude of differential gene expression along *Salix purpurea* chr15 **(D)** is portrayed as the product of the -$\log_{10}(p)$ and absolute $\log_2$(m/f) values. Lines connect orthologue pairs along *P. trichocarpa* (left) and *S. purpurea* (right) chr15 assemblies. For panels **(C)** and **(D)**, teal points represent genes considered differentially expressed (False Discovery Rate <0.05). The SDR interval boxplot was derived from mapping sex QTL within the F_2 *S. purpurea* family (n=497).

There were two primary gene clusters on *S. purpurea* chr15 that were significantly enriched for sex-biased expression span chr15:11.4-12.3 Mb and chr15:13.6-14.6 Mb (Figure 4.3D); the latter with more sex-biased genes and with higher significance than the former. Differentially-expressed genes located on unplaced *Salix* scaffolds 265, 582, 830, 1765, 2212, 2504, and 4349 align to *S. purpurea* chr15 gene clusters and corresponding regions on *Populus* chr15. For those genes located on unplaced *S. purpurea* scaffolds and chr15 which have high identity to *Populus* chr15 homologs, when ordered according to *Populus* chr15 positions, a substantial number of those genes appear to be duplicated within one or the other cluster.

Sex determination regions of dioecious species are often highly polymorphic, so we investigated the presence and absence of gene expression specific to F_1 females or males. Complete absence of expression in males was observed for *REM1* (SapurV1A.0301s0070), reticulon *RTNLB9* (SapurV1A.0530s0130), peptidase *M50B* (SapurV1A.0475s0170), chaperonin *TCP-1* (SapurV1A.1538s0020), *PMEI1* (SapurV1A.1210s0090), a gar2-related 18S pre-ribosomal assembly protein (SapurV1A.0178s0110), terpene synthase *TPS21* (SapurV1A.1522s0030), and *AGL98* (SapurV1A.2504s0020); all of which are located within or align to the pericentromeric SDR on *S. purpurea* chr15. Complete loss or very low-levels of gene expression in females was accompanied by correspondingly low-levels in males; nearly 10% of gene models were filtered from analyses for this reason.

Sex-biased genes highly expressed in females were enriched for GO terms in the biological processes of signaling, signal transmission and transduction, cation binding, and ion binding, as well as the molecular functions of copper ion binding, magnesium ion binding, signal transducer activity, and lyase activity. Genes showing higher expression in males were enriched for cell death, death, apoptosis, and programmed cell death, the molecular functions of ATP-

binding, structural constituent of the ribosome, and structural molecule activity, and the cellular components of intracellular organelle, ribonucleoprotein complex, and ribosome.

4.4.6 Sexually Dimorphic Inheritance of Gene Expression

Although I was not able to identify sex-biased expression in the F_1 internode transcriptome, considerable variation within the shoot tip transcriptome offered us a unique opportunity to dissect the heritable components of sexually dimorphic patterns of expression. Genes considered to exhibit sexually dimorphic inheritance were only reported for those with a significant nonadditive or transgressive inheritance class for at least one sex. Genes with expression inheritance classifications that did not show significant sex dimorphism were classified as having same-sex inheritance. Although a majority of the shoot tip transcriptome retained same-sex inheritance, 3.8% (1055 genes) displayed sexually dimorphic patterns of inheritance (Figure 5.8, Table 5.9).

While there were no significant differences in the median expression level for genes with same-sex inheritance, the expression levels of dimorphic genes were significantly greater in females than in males (Figure 5.8A). Broadly, sexually dimorphic inheritance in the F_1 shoot tip transcriptome was associated with a greater number of genes with conserved expression in males (65%) (Figure 5.8B) and nonadditive expression in females (75%) (Figure 5.8C). In addition, for those genes with sexually dimorphic inheritance, there was a marginally greater number of P2-dominant genes (174) in males compared to P1-dominant genes (148), whereas the opposite was found in females. Nearly five-times the number of dimorphic genes in females were classified as P1-dominant (499) compared with P2-dominant (119).

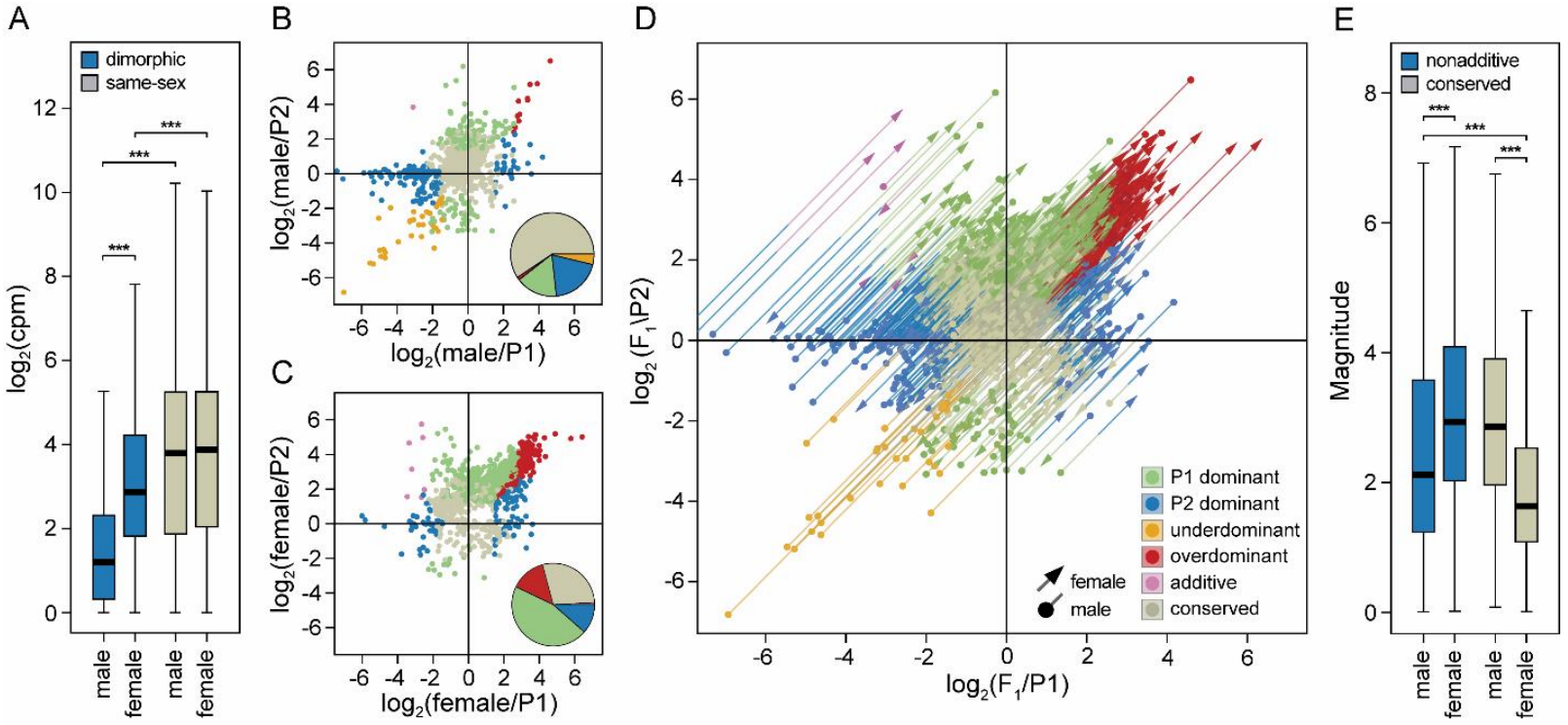

Figure 5.8 Sexually dimorphic inheritance in the F_1 shoot tip transcriptome. Boxplots **(A)** summarize the $\log_2$ normalized expression differences for genes with sexually dimorphic inheritance patterns (teal) and those with same-sex inheritance (beige), by sex. Asterisks above boxplots represent significant differences (Wilcoxon P<0.001). Scatterplots compare $\log_2$ normalized expression of F_1 males **(B)** and females **(C)** to the maternal (P1, x-axis) and paternal (P2, y-axis) expression. Points represent only genes with dimorphic inheritance patterns (same-sex inheritance not shown). Pie charts within the scatterplots summarize patterns of gene expression inheritance for genes with dimorphic gene expression for each sex. The scatterplot **(D)** illustrates overlain coordinates of gene expression inheritance for males and females, where each gene is represented by two vectors, one male (m_{xy}, points) and one female (f_{xy}, arrows), connected by a single line segment. Each segment is equally divided by two colors which correspond to the male and female inheritance class for each gene. The magnitude of dimorphic gene expression inheritance was calculated for each gene as the absolute Euclidean distance between the vectors, m_{xy} and f_{xy}, on the same Cartesian plane. For those genes with dimorphic inheritance, boxplot distributions **(E)** of nonadditive (blue) and conserved (beige) inheritance patterns for males and females depict differences in their absolute magnitude.

Table 5.9 Genes with sexually dimorphic inheritance patterns (females in rows and males in columns) and the total number of genes for dimorphic and same-sex gene expression inheritance classifications.

Inheritance	**P1 dominant**	**P2 dominant**	**Over-dominant**	**Under-dominant**	**Additive**	**Conserved**
females			*males*			
P1 dominant	0	66	2	2	1	428
P2 dominant	26	0	0	0	0	93
Overdominant	1	7	0	0	0	167
Underdominant	0	0	0	0	0	0
Additive	2	5	0	0	0	0
Conserved	119	96	9	31	0	0
overall						
male	322	217	19	33	1	27408
female	673	162	183	0	7	26975
dimorphic						
male	148	174	11	33	1	688
female	499	119	175	0	7	255
same-sex						
male and female	174	43	8	0	0	26720

The most drastic instances of sex dimorphism were for genes with transgressive (overdominant and underdominant) expression inheritance. Of the 175 overdominant genes in females, 167 had conserved expression in males, with only eight showing same-sex inheritance (Table 5.9). Further, of the 33 genes with an underdominant mode of expression inheritance, all were restricted to males. Relative to the results of the pooled F_1 family inheritance classifications (Table 5.4), additive inheritance plays a negligible role in both the female (7) and male (1) shoot tip transcriptome.

In order to quantify the extent of dimorphism for each gene, the magnitude of dimorphic expression was calculated as the Euclidean distance between male and female vectors on the same Cartesian plane of expression levels (Figure 5.9).

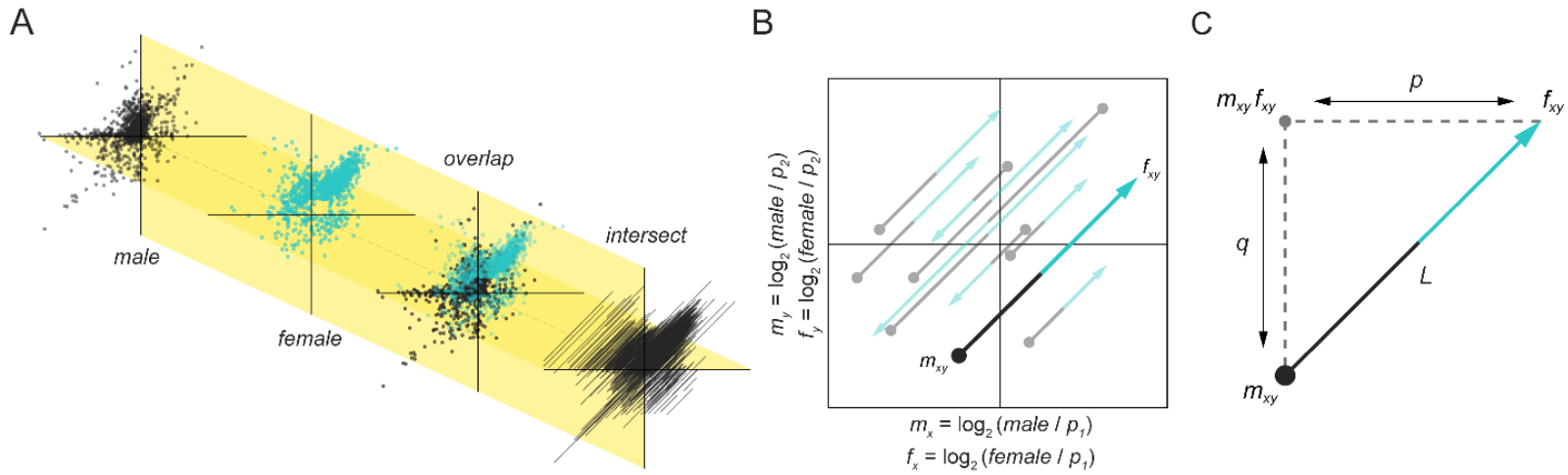

Figure. 5.9 Magnitude of sexually dimorphic gene expression inheritance. The scatterplot **(A)** demonstrates overlain Cartesian coordinates of gene expression inheritance for male and female individuals, where each gene is represented by two vectors, one male (m_{xy}, points) and one female (f_{xy}, arrowheads), connected by a single line segment. The absolute magnitude of dimorphic gene expression inheritance **(B)** was determined as **(C)** the square root of the sum of the lengths, p^2 and q^2, or the Euclidean distance (L^2) between vectors m_{xy} and f_{xy}, such that: $L^2 = \langle p, q \rangle = [(|m_x| - |f_x|)^2 + (|m_y| - |f_y|)^2]^{0.5}$, where m_x is the male coordinate derived from $\log_2(m/p_1)$, m_y is the male coordinate derived from $\log_2(m/p_2)$, f_x is the female coordinate derived from $\log_2(f/p_1)$, and f_y is the female coordinate derived from $\log_2(f/p_2)$, p is the squared absolute difference of the vectors $(|m_x|, |f_x|)$ and $(|m_x|, |m_y|)$, and q is the squared absolute difference of the vectors $(|m_x|, |f_y|)$ and $(|m_x|, |m_y|)$.

The direction and magnitude of dimorphic gene expression in the F_1 shoot tip transcriptome is illustrated in Figure 5.9D. The contribution of nonadditive inheritance on dimorphic gene expression was examined directly, because gene expression inheritance in an individual is simply a function of the difference in progeny and parent expression ratios. This method demonstrates that there was a positive relationship between the magnitude of dimorphic inheritance and midparent differential expression. In F_1 males, sexually dimorphic genes with conserved inheritance exhibited significantly greater differences in the magnitude of differential expression than genes with nonadditive expression. However, dimorphic genes with conserved inheritance in females (and nonadditive inheritance patterns in males) exhibited a significantly lower magnitude of expression than genes with nonadditive expression (Figure 5.9E).

Likewise, the absolute magnitude of genes showing conserved expression in males was not significantly different from the magnitude of nonadditive expression in females, signifying that a higher frequency of sexually dimorphic genes with conserved inheritance in males coincided with nonadditive gene expression in females. Simply, as the distance between female and male vectors (magnitude) increased, the frequency of nonadditive and conserved patterns of inheritance increased, respectively.

5.4.6 Sexually Dimorphic ASE

To test the hypothesis that sex dimorphic gene expression was attributable differences in regulation, I compared the magnitude of *cis*- and *trans*-regulation of genes expressed predominantly in either females or males. In order to determine the extent to which patterns of regulatory divergent expression varies for the same gene, I contrasted the magnitude and direction of regulatory divergence and the proportion of their effects, based solely on sex in the F_1. Sexually dimorphic ASE was considered to be the total number of genes that significantly

differ by regulatory class between three F_1 male individuals and three F_1 female individuals. Genes assigned regulatory classes which did not differ between males and females were characterized as having same-sex ASE. For each gene showing significant regulatory divergence in either sex, these results illustrate the total number of genes that show contrasting or sexually dimorphic regulatory patterns. Only genes from the ASE analysis which were unique to either females or males were considered.

In the shoot tip transcriptome, a total of 297 genes displayed sexually dimorphic ASE (Table 5.10, Figure 5.10), where 143 had higher expression levels in F_1 males and 154 were expressed greater in F_1 females. For genes showing sexually dimorphic ASE, conserved or *cis*-regulation in females coincided with *cis* × *trans* or compensatory regulation for the same genes in males. For instance, there were significantly greater numbers of dimorphic genes expressed in the male internode transcriptome, with 54 *cis* × *trans* and 88 compensatory regulated genes, compared to only 4 *cis* × *trans* and 9 compensatory regulated genes uniquely expressed in females (Table 5.10). By comparing ASE of female and male F_1 progeny individuals, genes with significant *cis*-regulation were significantly biased towards maternal parent (P1) in the shoot tip and internode transcriptome of females, as well as the shoot tip transcriptome of males, irrespective of the magnitude of ASE (Table 5.10). On a per-site basis, antagonistic *cis* × *trans* and compensatory regulation in the F_1 shoot tip transcriptome was driven by upregulation of the paternal (P2) allele, whereas the maternal (P1) allele was upregulated for either pure *cis*- or *trans*-regulation or *cis* + *trans* regulation (Figure 5.10B).

The absolute magnitude of *trans*-regulated genes was greater than that of *cis*-regulated genes for both males and females (Figure 5.10C). Although the magnitude of effect class sizes of genes that showed pure *cis*- or *trans*-regulation were not significantly different by sex, a

significantly greater frequency of *cis*-effects were present at the tails (< 0.5 and > twofold) and a greater frequency of *trans*-effects between tails (> 0.5 and < twofold) (Figure 5.10D). By scaling *cis*- and *trans*-regulatory divergence by the total regulatory divergence, as in Coolon et al., (2013), I obtained a relative percent divergence due to *cis*- and *trans*-effects. With regard to females, the proportion of genes with evidence of *cis*-regulatory divergence was significantly greater than was observed in males. The proportion of genes showing significant *trans*-regulatory divergence was greater in females than in males, but with weaker significance than the *cis*-regulated genes (Figure 5.10E).

Table 5.10 Sexually dimorphic patterns of regulatory divergence.

	cis	***trans***	***cis* + *trans***	***cis* × *trans***	**compensatory**	**ambiguous**	**conserved**
Internode							
dimorphic							
males	67	60	2	**54**	**88**	142	**39**
females	46	52	2	**4**	**9**	171	**168**
same-sex	50	62	2	17	22	698	4008
Shoot tip							
dimorphic							
males	23	35	2	17	14	**162**	45
females	46	42	3	16	12	**89**	89
same-sex	29	31	0	6	3	757	3857

Dimorphic ASE was considered as genes within regulatory divergence classifications (FDR=0.005) that differ between males and females, and same-sex ASE were those with the same regulatory classifications. Numbers in boldface indicate significant (P<0.01) deviations from a 1:1 ratio according to a χ^2 test.

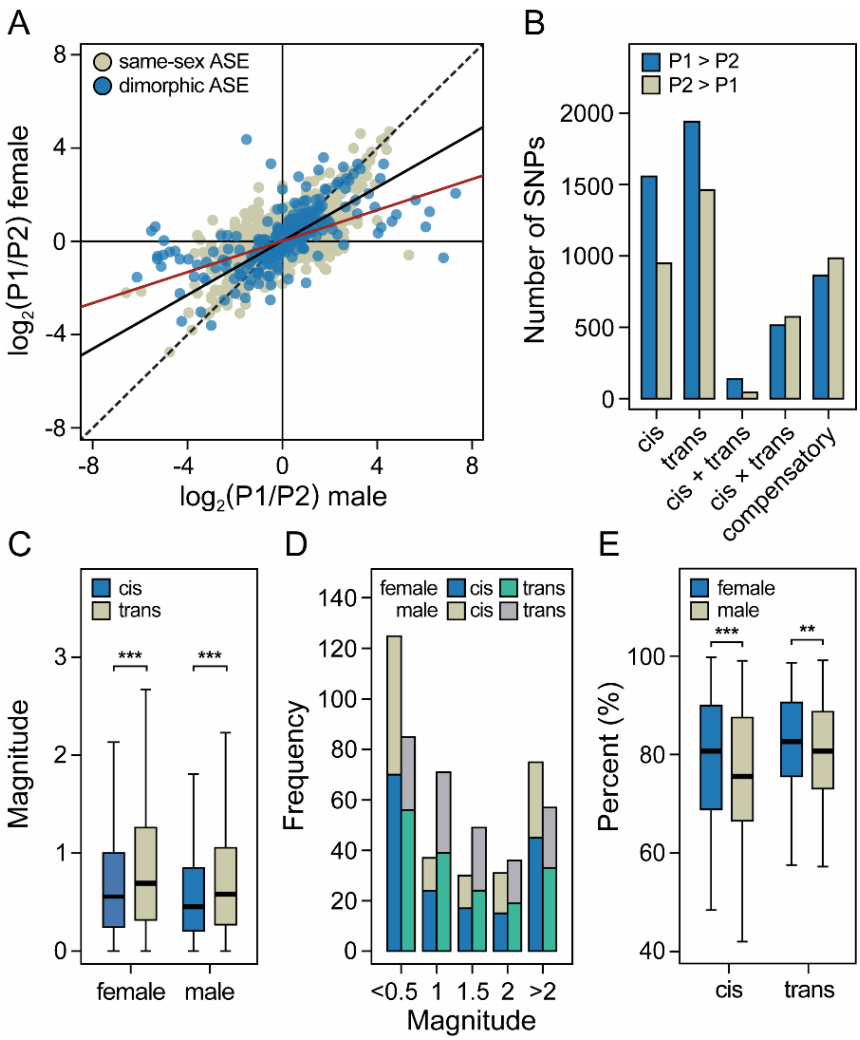

Figure 5.10 Sexually dimorphic regulatory divergence in the F_1 shoot tip transcriptome. The ratio of expression of parent alleles **(A)** is plotted for females (y-axis) versus males (x-axis) showing dimorphic ASE (blue points) and same-sex ASE (beige points) of the shoot tip transcriptome. The solid red line and solid black line represent the slopes of dimorphic ASE and same-sex ASE, respectively, where the black dashed line has an intercept=0 and slope=1. The barplots **(B)** explore the cases of parental dominance in hybrid regulatory patterns. Boxplots **(C)** show the distribution of the absolute differences between males and females for *cis* (blue) and *trans* (beige) for genes showing dimorphic ASE. Asterisks ***, ** above boxplots denote significant differences at a Wilcoxon $P < 0.001$ and $P < 0.01$, respectively. Barplot **(D)** distributions of genes with significant *cis*- and *trans*-regulation were binned according to the magnitude of their effect class size for females (blue and teal, respectively) and males (beige and grey, respectively). Boxplots **(E)** summarize the percent divergence due to *cis*- and *trans*-effects, where *cis* = $\log_2$(P1H / P2H) and *trans* = $\log_2$(P1H / P2H) - $\log_2$(P1 / P2), and *% cis* = [|*cis*| / (|*cis*| + |*trans*|)] × 100 and *% trans* = [|*trans*| / (|*cis*| + |*trans*|)] × 100.

Table 5.11 Number of pure *cis*- and *trans*-regulated genes in male and female tissues.

	Shoot tip				**Internode**			
	cis		***trans***		***cis***		***trans***	
	P1	**P2**	**P1**	**P2**	**P1**	**P2**	**P1**	**P2**
Males								
All genes	**40**	**16**	37	34	66	51	68	54
>1.5-fold	**31**	**10**	1	4	48	36	2	5
>2-fold	**21**	**7**	1	1	34	29	0	4
Females								
All genes	**64**	**25**	46	36	**79**	**37**	71	50
>1.5-fold	**47**	**10**	1	2	**59**	**23**	3	1
>2-fold	**38**	**5**	1	1	**49**	**18**	0	1

Numbers in boldface indicate significant ($P<0.01$) deviations from a 1:1 ratio according to a χ^2 test.

5.5 Discussion

5.5.1 Dominance and Regulatory Divergence in F_1 and F_2 S. purpurea

The majority of gene expression inheritance and ASE studies in both plants and animals have focused on F_1 hybrids generated from crossing stable, inbred parents. Unlike what has been described in model crop plants, like maize (Stupar and Springer, 2006) and rice (Song et al., 2013), I show that in *S. purpurea*, the greatest proportion of differentially expressed genes did not exhibit a primarily additive mode of inheritance, but rather showed strong patterns of dominance. Preference of uniparental expression in progeny is thought to be orchestrated by epistatic interactions, which primarily function to silence one of the parental alleles in a parent-of-origin manner (Chen and Pikaard, 1997; Stupar et al., 2007). Here, maternal dominance represented the greater proportion of the nonadditive gene expression in both tissue transcriptomes of the F_1 and F_2 families. Other cases of expression-level dominance has been described in hybrids of intraspecific thistle (Bell et al., 2013), interspecific coffee (Combes et al., 2015), synthetic allotetraploid rice (Xu et al., 2014), as well as allotetraploid Arabidopsis (Shi et al., 2012). Reciprocal hybridization has become a useful technique to examine genomic imprinting by comparing common patterns of uniparental expression of alleles in reciprocal family progeny (Donoghue et al., 2014; Baldauf et al., 2016). In developing seeds of Arabidopsis, there is strong evidence that imprinting genes regulate early endosperm development and nutrient translocation and partitioning in the seed (McKeown et al., 2011). While epistasis may well contribute to the prodigious levels of dominant gene expression observed in F_1 and F_2 *S. purpurea*, since *Salix* spp. are dioecious, reciprocal crosses cannot be generated in the classical sense.

The inheritance and regulatory patterns described in this study were most similar to what was reported in hybrids derived from heterozygous parents that were collected from natural *C. arvense* populations (Bell et al., 2013). For instance, for both F_1 *S. purpurea* and F_1 *C. arvense*, more divergently expressed genes showed higher expression of the maternal allele than the paternal allele and that a significantly greater proportion of dominant cases had maternal expression patterns. Importantly, there were similar trends between *S. purpurea* and *C. arvense* in *cis*- and *trans*-divergence, such that *trans*-divergence was associated with higher expression of the paternal allele, whereas *cis*-divergence tended to increase expression of the maternal allele. However, there were differences in transgressive inheritance classes between the species. The frequency of overdominant gene expression in *S. purpurea* (shoot tip) was significantly less than underdominant expression, whereas the opposite was detected in *C. arvense*.

The transcription of two divergent parental alleles in the hybrid can be controlled by *cis*-regulatory elements as well as *trans*-acting factors, but parental expression differences may be reduced in the hybrid for genes subject to strong *trans*-regulation (Xu et al., 2014). Allele-specific tests in hybrids of inbred maize (Stupar and Springer, 2006) established that *cis*-acting regulatory variation accounted for the majority of the observed parental expression divergence and that pure *cis*-regulation correlated with additive expression patterns in the F_1 hybrid. In synthetic allotetraploids of *indica* and *japonica* rice subspecies, parental expression differences were found to be intensified in the allotetraploid but showed a reduction of expression divergence in reciprocal F_1 hybrids, owning to the effects of a common *trans* environment on divergent *cis*-factors (Xu et al., 2014). Using allotetraploids of *Arabidopsis thaliana* as well as resynthesized allotetraploids of *A. arenosa*, Shi et al. (2012) showed that the higher level of sequence divergence (and dominance) in *A. arenosa* promote flexibility of *trans*-factors for their

binding to interacting factors and *cis*-elements of *A. thaliana* and *A. arenosa* alleles. In F_1 and F_2 *S. purpurea*, both *cis*- and *trans*-effects account for nonadditive or transgressive gene expression. These data also show significantly greater *cis* × *trans* and *cis*-*trans*-compensatory regulation in the F_1 shoot tip transcriptome compared to the internode transcriptome, fitting the model in which greater transgressive inheritance tends to show greater proportions of *cis* × *trans* regulatory divergence (McManus et al., 2010). Further, for genes showing both *cis*- and *trans*-effects, antagonistic *cis* × *trans* and compensatory interactions were driven by the up-regulation of the paternal allele. The compatibility of novel *trans*-factors with their target binding sites could explain the high proportion of expression-level dominance in *S. purpurea*, which suggests that compensatory interactions and stabilizing selection may play an important role in maintaining parental gene expression levels.

5.5.2 The Fasciclin Gene Family is Highly Expressed in the F_1 Internode

Fasciclin-like AGPs were found to be over-represented and highly expressed in stem internode tissue, whereas genes encoding defense-related proteins were primarily upregulated in shoot tip tissue. Found in most angiosperms, AGPs are a class of Hyp-rich glycoproteins that are highly abundant in the cell wall and plasma membrane which are involved in various aspects of plant growth (Kitazawa et al., 2013) and primarily function in cell adhesion (Johnson et al., 2003). In *Populus* and *Eucalyptus*, FLAs were highly expressed during xylem differentiation (Hefer et al., 2015) and implicated in secondary cell wall thickening (Lafarguette et al., 2004). *Salix FLA11* and *FLA12* gene models are closely related to Arabidopsis *FLA11* and *FLA12* homologs and have been shown to contribute to stem tensile strength and the modulus of stem elasticity (MacMillan et al., 2010; MacMillan et al., 2015), which is consistent with FLA roles in cellulose deposition during secondary xylogenesis. These data suggest that constitutive defense

in the shoot tips of *S. purpurea* is coordinated with the rapid development of secondary xylem in stem internodes. Functional analysis of FLA homologs in *Salix* may prove to be useful in characterizing genes involved in the regulation of stem strength for the genetic improvement of shrub willow bioenergy crops.

5.5.3 Sexually Dimorphic Gene Expression

The effect of sex may explain a significant amount of the variation driving the evolution of gene expression in dioecious plants. While differential expression in the F_1 family shoot tip transcriptome was found to be almost entirely nonadditive, the primary contributors of nonadditive expression were only discernable by barcoding and sequencing the F_1 progeny individuals. This study provides evidence of sexually dimorphic expression in intraspecific F_1 *S. purpurea*, where both *cis*- and *trans*-effects accounted for the observed differences in magnitude among regulatory patterns; however, the effect of sex on gene expression in the F_1 was tissue-specific. Although allelic effects were comparable, Mieklejohn et al. (2014) described sexually dimorphic regulatory divergence among *Drosophilia simulans* and *D. mauritiana* introgression hybrids, proposing that pure-species genotypes carry modifier alleles that increase sexually dimorphic expression.

While *cis*-effects accounted for more of the regulatory divergent expression in F_1 females, compensatory regulation was enriched in F_1 males. It may be the sex determining system itself can help explain sexually dimorphic ASE in *S. purpurea*. In females, the Z haplotype is paternally (ZZ) inherited and the W is maternally (ZW) inherited, but in males, both the maternal and paternal Z haplotypes are inherited equally. If there is substantial divergence between the Z and W, theoretically, *cis*-effects should outweigh *trans*-effects in females. In contrast, the two Z

haplotypes present in males are considerably less polymorphic, such that parental expression differences on the Z would be explained by more antagonistic *trans*-regulatory interactions.

It is possible that a subset of the sexually dimorphic genes in F_1 *S. purpurea* actually reflect unresolved conflicts between females and males and that sex-biased gene expression on autosomes or in pseudo-autosomal regions could be due to sexual antagonism (Alström-Rapaport et al., 1997; Nagamitsu and Futamura, 2014; Su et al., 2016; Zhai et al., 2016). The accumulation of sexually-antagonistic loci are predicted to occur in newly formed SDRs, where differential selection via tight linkage is beneficial to one sex and harmful to the other (Rice 1992). Over time, accumulation of sexually antagonistic genes within a population may eventually lead to a considerable conflict between sexes such that adaptation by each sex would be compromised. As female-benefit and male-detriment genes accumulated, Rice (1992) found that the sex ratios (m/f) of Drosophilia declined, suggesting that pseudo-autosomal regions (PARs) near the SDR can act as a hot spot for the accumulation of genes detrimental to the homogametic sex.

There are a number of studies which have reported consistent 2:1 (f:m) sex ratio-biases in natural willow populations (Ueno et al., 2007; Che-Castaldo et al., 2015). Disproportionate sex-ratios in natural populations of *Salix* spp. could promote evolutionary biases by sustaining the regulatory roles of the predominant sex. The higher proportion of genes with female-specific expression in *S. purpurea* may be a consequence of cyclic asexual reproduction leading to a relaxation of purifying selection on male-biased genes, whereby conflicting modes of inheritance and regulatory divergence patterns could lead to unequal investments in reproduction and reproductive strategies (Parsch and Ellegren, 2013). However, sex-biased gene expression in *S. purpurea* may reflect resolved conflicts in favor of females.

5.5.4 Sex-Biased Expression Localizes to the SDR of S. purpurea chr15

The genus *Salix* exhibits a ZW sex determination system (Alström-Rapaport et al., 1997; Gunter et al., 2003; Semerikov et al., 2003; Liu et al., 2013), where the female is heterogametic (ZW) and the male is homogametic (ZZ), in contrast to the male heterogametic XY system of *P. trichocarpa* (Yin et al., 2008; Geraldes et al., 2015). Although the genomes are rather collinear, is not entirely clear whether the SDR developed before or after the Salicoid duplication (Rodgers-Melnick et al., 2012) and divergence of *Salix* from *Populus* (Hou et al., 2016). I have delimited the SDR of *S. purpurea* to a centromeric region on chr15 (Zhou et al., 2018) using a full-sib F_2 *S. purpurea* mapping population (described here) and a diverse panel of *S. purpurea* naturalized North American genotypes. Zhou et al. (2018) show in *S. purpurea* that females are heterozygous (or hemizygous) and are males homozygous at polymorphic sites within the coding sequences of genes found in the SDR of *S. purpurea* chr15. As low recombination near the *Salix* SDR has prevented the identification of a single gene responsible for sex determination in experimental mapping populations (Pucholt et al., 2015; Chen et al., 2016; Pucholt et al., 2017), increasing the overall experimental scale is likely required to pin-point the causative sex-determining genes.

Of the previous gene expression studies on sex determination in *Salix*, none have yet identified the gene(s) responsible. Even though a relatively small proportion of genes in the F_1 shoot tip transcriptome of *S. purpurea* are sex-biased (< 0.1%), there was no substantial evidence of sex-biased expression in the F_1 internode transcriptome. At the time shoot tips were collected, it was likely that floral buds were developing near the shoot apical meristem and it is possible this amalgam of cells included a subset of that were expressing genes involved in the determination of male or female flowers. Genes differentially expressed among the F_1 male and

female shoot tip transcriptome are not necessarily indicative of sex determination alone, but altogether represent a cascade of developmental gene expression involved in patterning, signaling, and organ suppression during the vegetative-to-reproductive transition leading to sexual dimorphism (Fairbairn and Roff, 2006).

Tests for differential expression between mature catkins of female and male willows likely confound the search for sex determining genes because there are profound morphological and phenological differences between females and males at floral maturity. For instance, RNA-sequencing of male and female catkins of *S. suchowensis* (Liu et al., 2013) and *S. viminalis* (Pucholt et al., 2017) identified a plethora of differentially-expressed genes, but failed to establish any biological links among those with significant associations. Yet, a link between sex determination and meristem fate is well-described in oil palm (Ho et al., 2016) and in maize via RNA-induced silencing of *TS1* and *AP2* by miRNA172 (Hartwig, 2011). Genes related to DNA methylation, *MET1* and *DDM1*, as well as SAUR-like auxin responsive genes were implicated in sex determination in andromonecious *Populus tomentosa* (Song et al., 2013) and *P. trichocarpa* (McKown et al., 2017). In *Asparagus officinalis*, the MYB-like transcription factor, *MSE1*, is specifically expressed in males and has been shown to induce male sterility in knockouts of Arabidopsis (Murase et al., 2017). The identification of *MeGI*, an autosomal homeobox transcription factor in *Diospyros lotus*, has been shown to dominantly suppress male organ development can be suppressed by the small RNA *OGI* on the Y chromosome that targets *MeGI* for gene silencing (Akagi et al., 2014).

The most significant sex-biased gene highly expressed in males ($-\log_{10}(p)=6.4$) was most similar to the 40S ribosomal subunit, *RPS3* (SapurV1A.0582s0010), a positive regulator of apoptosis during UV irradiation in Arabidopsis (Lee et al., 2010). Conversely, the most

significant upregulated gene in females ($-\log_{10}(p)=42.6$) was most similar to an inhibitor of apoptosis in Arabidopsis, a C3CH zinc finger homolog of *NUCLEAR INTERACTING PARTNER OF ALK* (*NIPA*) (SapurV1A.0934s0010). Developmental requirements controlling the vegetative-to-reproductive transition in meristematic tissues are likely to differ by sex. The Arabidopsis *ASSOCIATE OF C-MYC* (*AMY1*) homolog in *S. purpurea* (SapurV1A.1254s0040), was highly expressed in females and has been shown to interact with numerous chromatin modifiers and transcription factors in Arabidopsis (Taira et al., 1998) and implicated as a universal amplifier of gene expression, acting to increase output at all active promoters. Programmed cell death (PCD) in developing floral buds could lead to the specification of sex at the cost of structural reproductive components integral to the opposite sex. If PCD is a primary component in the determination of male or female flowers in *S. purpurea*, both the greater numbers and magnitude of differential gene expression in females could indicate that the role of PCD is more active in the development of female flowers.

REPRODUCTIVE MERISTEM 1 (*REM1*) (SapurV1A.0301s0070), *AGAMOUS-LIKE 98* (AGL98) (SapurV1A.2504s0020), and *DOWN-REGULATOR OF TRANSCRIPTION 1* (*DR1*) (SapurV1A.0301s0160) were highly expressed in F_1 female shoot tip tissues and among the top differentially expressed genes that encode for known flowering-time genes in Arabidopsis (Pagnussat et al., 2005). The *REM1* gene aligns to a region within the SDR of *S. purpurea* chr15 and highly expressed in females, but showed complete null expression in males. In maturing inflorescences of Arabidopsis, *REM1* expression localized to only a few vegetative cells in the shoot apical meristem, but during the vegetative-to-reproductive transition, *REM1* was progressively restricted to the gynoecium which gives rise to the stigma, style, and septum (Franco-Zorrilla et al., 2002). Although SapurV1A.1250s0040 is paralogous to

SapurV1A.0301s0070, the former is highly expressed in males and is homologous to *A. thaliana AGO9*. Implicated in the vegetative-to-reproductive transition during male gametogenesis in *A. thaliana*, the primary role of *AGO9* is to silence transposable elements (TEs) in the female gametophyte; thereby establishing the transgenerational epigenetic information required to control gametophytic fate (Hernandez-Lagana et al., 2016). Thus, it is conceivable that there is sRNA-induced silencing of W-specific genes via AGO loading in the RdRM pathway.

I have identified *TCP-1* (SapurV1A.1538s0020) as one of the few sex-biased genes that exhibited complete null expression in male shoot tip. TCP transcription factors play pivotal roles in the control of shoot morphogenesis by negatively regulating the expression of boundary-specific genes (Koyama et al., 2007; Koyama et al., 2010; Li et al., 2012), such as suppression of secondary wall thickening of the anther endothecium in Arabidopsis (Wang et al., 2015). Geraldes et al. (2015) identified two *TCP-1* chaperonin family cpn-60 proteins associated with sex determination in the T52 *P. trichocarpa* association population, and this gene could play a major role in suppression of the anther endothecium in female *S. purpurea* catkins. Two sex-biased genes encoding the DExH-box ATP-dependent RNA helicase, *BRR2C,* are highly conserved components of the spliceosome and are required for efficient splicing of *FLC* introns, as well as regulation of *FT* and *SOC1* in Arabidopsis (Mahrez et al., 2016). Of the two sex-biased Dr1/NF-Y paralogs in *S. purpurea*, SapurV1A.0934s0060 was highly expressed in males, whereas SapurV1A.0301s0160 was highly expressed in females. Dr1 represses RNAP II transcription by binding to TBP to prevent the formation of an active transcription complex. Members of the heterotrimeric NF-Y transcription factor family in Arabidopsis initiate photoperiod-dependent flowering and also required for activation of the *FT* promoter by initiating downstream events leading to floral transition (Siriwardana et al., 2016).

I observed a number of genes with null expression in males that are highly expressed in females, yet we found no evidence for the converse. Rather, low levels of gene expression in females was always accompanied by low expression levels in males. Nearly 65% of all sex-biased genes were more highly expressed in females than in males, indicating disproportionate W-specific genes or Z-specific pseudogenes in the SDR. Given our findings on sexually dimorphic expression in *S. purpurea*, a reasonable hypothesis is that the transcription of genes involved in early floral meristem identity is differentially regulated by the relative abundance of RNAP II core subunits, whose promoter site specificity, initially guided by TBP components, are likely sex-specific.

I conclude that chr15 contributes to sexual dimorphism in *S. purpurea*, as genes with sex-biased expression were vastly over-represented within or near the SDR, compared to other autosomes. Although there was no apparent localization of sex-biased genes along chr19 as was found for chr15, over 12% of the sex-biased genes were most similar to *P. trichocarpa* chr19 gene models, many of which were highly expressed in males. In addition to mapping sex QTL on all three linkage maps to chr15 in F_2 *S. purpurea* (family 317), Zhou et al. (2018) identified a secondary sex QTL on chr19; however, this QTL was only present in the male backcross map. It is not clear whether chr19 is epistatic to chr15, yet it seems likely that chr19 is the ancestral sex determining chromosome of *Salix* and *Populus*, and may very well continue to contribute to sex determination, sex ratio bias, and sex dimorphism in *Salix*.

This study provides the first detailed analysis of transcriptome-wide regulatory divergent expression in *Salix*. Expression-level dominance and sexual dimorphism are prevailing features of differential gene expression in *S. purpurea*. Expanding upon transcriptomic resources in *Salix* will not only contribute to our understanding of the evolution of dioecy in the Salicaceae, but

also facilitate the functional characterization of genes underlying sex determination in dioecious species.

5.6 REFERENCES

Akagi T, Henry IM, Tao R, Comai L. 2014. A Y-chromosome-encoded small RNA acts as a sex determinant in persimmons. *Science*, 346: 646–650.

Alström-Rapaport C, Lascoux M, Gullberg U. 1997. Sex determination and sex ratio in the dioecious shrub *Salix viminalis* L. *Theoretical and Applied Genetics*, 94: 493–497.

Argus GW. 1997. Infrageneric classification of *Salix* (*Salicaceae*) in the New World. *Monographs in Systematic Botany*, 52: 1–121.

Baldauf JA, Marcon C, Paschold A, Hochholdinger F. 2016. Nonsyntenic genes drive tissue-specific dynamics of differential, nonadditive, and allelic expression patterns in maize hybrids. *Plant Physiology*, 171: 1144–1155.

Bell GD, Kane NC, Rieseberg LH, Adams KL. 2013. RNA-seq analysis of allele-specific expression, hybrid effects, and regulatory divergence in hybrids compared with their parents from natural populations. *Genome Biology and Evolution*, 5: 1309–1323.

Berlin S, Lagercrantz U, von Arnold S, Ost T, Ronnberg-Wastljung A. 2010. High-density linkage mapping and evolution of paralogs and orthologs in *Salix* and *Populus*. *BMC Genomics*, 11: 129.

Cameron KD, Phillips IS, Kopp RF, Volk TA, Maynard CA, Abrahamson LP, Smart LB. 2008. Quantitative genetics of traits indicative of biomass production and heterosis in 34 full-sib F_1 *Salix eriocephala* families. *BioEnergy Research*, 1: 80–90.

Carlson CH, Smart LB. 2016. Electrical capacitance as a predictor of root dry weight in shrub willow (*Salix*; Salicaceae) parents and progeny. *Applications in Plant Sciences*, 4: 1600031.

Chen CZ, Pikaard CS. 1997. Transcriptional analysis of nucleolar dominance in polyploid plants: Biased expression/silencing of progenitor rRNA genes is developmentally regulated in *Brassica*. *Proceedings of the National Acadamy of Sciences of the United States of America*, 94: 3442–3447.

Chen Y, Wang T, Fang L, Li X, Yin T. 2016. Confirmation of single-locus sex determination and female heterogamety in willow based on linkage analysis. *PLoS ONE,* 11: e0147671.

Coate JE, Doyle JJ. 2010. Quantifying whole transcriptome size, a prerequisite for understanding transcriptome evolution across species: an example from a plant allopolyploid. *Genome Biology and Evolution,* 2: 534–546.

Combes MC, Hueber Y, Dereeper A, Rialle S, Herrera JC, Lashermes P. 2015. Regulatory divergence between parental alleles determines gene expression patterns in hybrids. *Molecular Biology and Evolution,* 7: 1110–1121.

Coolon JD, Webb W, Wittkopp PJ. 2013. Sex-specific effects of *cis*-regulatory variants in *Drosophila melanogaster*. *Genetics,* 195: 1419–1422.

DePristo MA, Banks E, Poplin RE, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics,* 43: 491–498.

Donoghue MT, Fort A, Clifton R, Zhang X, McKeown PC, Voight-Zielinski ML, Borevitz JO, Spillane C. 2014. C(m)CGG methylation-independent parent-of-origin effects on genome-wide transcript levels in isogenic reciprocal F1 triploid plants. *DNA Research,* 21: 141–151.

Du Z, Zhou X, Ling Y, Zhang Z, Su Z. 2010. agriGO: a GO analysis toolkit for the agricultural community. *Nucleic Acids Research*, 38: W64–W70.

Fabio ES, Kemanian AR, Montes F, Miller RO, Smart LB. 2017. A mixed model approach for evaluating yield improvements in interspecific hybrids of shrub willow, a dedicated bioenergy crop. *Industrial Crops and Production*, 96: 57–70.

Fabio ES, Volk TA, Miller RO, Serapiglia MJ, Gauch HG, Van Rees KCJ, Hangs RD, Amichev BY, Kuzovkina YA, Labrecque M, Johnson GA, Ewy RG, Kling GJ, Smart LB. 2016. Genotype × environment interactions analysis of North American shrub willow yield trials confirms superior performance of triploid hybrids. *GCB Bioenergy*, 9: 445–459.

Fairbairn DJ, Roff DA. 2006. The quantitative genetics of sexual dimorphism: assessing the importance of sex-linkage. *Heredity*, 97: 319–328.

Franco-Zorrilla JM, Cubas P, Jarillo JA, Fernández-Calvín B, Salinas J, Martínez-Zapater JM. 2002. *AtREM1*, a member of a new family of B3 domain-containing genes, is preferentially expressed in reproductive meristems. *Plant Physiology*, 128: 418–427.

Geraldes A, Hefer CA, Capron A, Kolosova N, Martinez-Nuñez F, Soolanayakanahally RY, Stanton B, Guy RD, Mansfield SD, Douglas CJ, Cronk QC. 2015. Recent Y chromosome divergence despite ancient origin of dioecy in poplars (*Populus*). *Molecular Ecology*, 24: 3243–3256.

Gunter LE, Roberts GT, Lee K, Larimer FW, Tuskan GA. 2003. The development of two flanking SCAR markers linked to a sex determination locus in *Salix viminalis* L. *Journal of Heredity*, 94: 185–189.

Hanley SJ, Karp A. 2013. Genetic strategies for dissecting complex traits in biomass willows (*Salix* spp.). *Tree Physiology*, 34: 1167–1180.

Hartwig T, Fujioka S, Klempien A, Weizbauer R, Potluri DPV, Choe S, Johal GS, Schulz B. 2011. Brassinosteroid control of sex determination in maize. *Proceedings of the National Acadamy of Sciences of the United States of America*, 108: 19814–19819.

Hefer CA, Mizrachi E, Myburg AA, Douglas CJ, Mansfield SD. 2015. Comparative interrogation of the developing xylem transcriptomes of two wood-forming species: *Populus trichocarpa* and *Eucalyptus grandis*. *New Phytologist*, 206: 1391–1405.

Hernandez-Lagana E, Rodriguez-Leal D, Lua J, Vielle-Calzada JP. 2016. A multigenic network of ARGONAUTE4 clade members controls early megaspore formation in *Arabidopsis*. *Genetics*, 204: 1045–1056.

Ho H, Low JZ, Gudimella R, Tammi MT, Harikrishna JA. 2016. Expression patterns of inflorescence- and sex-specific transcripts in male and female inflorescences of African oil palm (*Elaeis guineensis*). *Annals of Applied Biology*, 168: 274–289.

Hou J, Ye N, Dong Z, Lu M, Li L, Yin T. 2016. Major chromosomal rearrangements distinguish willow and poplar after the ancestral "Salicoid" genome duplication. *Genome Biology and Evolution*, 8: 1868–1875.

Johnson KL, Jones BJ, Bacic A, Schultz CJ. 2003. The fasciclin-like arabinogalactan proteins of Arabidopsis. A multigene family of putative cell adhesion molecules. *Plant Physiology*, 133: 1911–1925.

Kitazawa K, Tryfona T, Yoshimi Y, Hayashi Y, Kawauchi S, Antonov L, Tanaka H, Takahashi T, Kaneko S, Dupree P, Tsumuraya Y, Kotake T. 2013. β-galactosyl Yariv reagent binds to the beta-1,3-galactan of arabinogalactan proteins. *Plant Physiology*, 161: 1117–1126.

Kopp RF, Smart LB, Maynard CA, Isebrands JG, Tuskan GA, Abrahamson LP. 2001. The development of improved willow clones for eastern North America. *The Forestry Chronocle*, 77: 287–292.

Koyama T, Furutani M, Tasaka M, Ohme-Takagi M. 2007. TCP transcription factors control the morphology of shoot lateral organs via negative regulation of the expression of boundary-specific genes in *Arabidopsis*. *The Plant Cell*, 19: 473–484.

Koyama T, Mitsuda N, Seki M, Shinozaki K, Ohme-Takagi M. 2010. TCP transcription factors regulate the activities of *ASYMMETRIC LEAVES1* and miR164, as well as the auxin response, during differentiation of leaves in *Arabidopsis*. *The Plant Cell*, 22: 3574–3588.

Lafarguette F, Leplé J-C, Déjardin A, Laurans F, Costa G, Lesage-Descauses M-C, Pilate G. 2004. Poplar genes encoding fasciclin-like arabinogalactan proteins are highly expressed in tension wood. *New Phytologist*, 164: 107–121.

Landry CR, Wittkopp PJ, Taubes CH, Ranz JM, Clark AG, Hartl DL. 2005. Compensatory *cis-trans* evolution and the dysregulation of gene expression in interspecific hybrids of *Drosophilia*. *Genetics*, 171: 1813–1822.

Lee SB, Kwon I-S, Park J, Lee H-K, Ahn Y, Lee C, Kim J, Choi SY, Cho S-W, Ahn J-Y. 2010. Ribosomal protein S3, a new substrate of Akt, serves as a signal mediator between neuronal apoptosis and DNA repair. *The Journal of Biological Chemistry*, 285: 29457–29468.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25: 1754–1760.

Li Z, Li B, Shen WH, Huang H, Dong A. 2012. TCP transcription factors interact with AS2 in the repression of class-I *KNOX* genes in *Arabidopsis thaliana*. *Plant Journal*, 71: 99–107.

Liu J, Yin T, Ye N, Chen Y, Yin T, Liu M, Hassani D. 2013. Transcriptome analysis of the differentially expressed genes in the male and female shrub willows (*Salix suchowensis*). *PLoS ONE*, 8: e60181.

MacMillan CP, Mansfield SD, Stachurski ZH, Evans R, Southerton SG. 2010. Fasciclin-like arabinogalactan proteins: specialization for stem biomechanics and cell wall architecture in Arabidopsis and *Eucalyptus*. *Plant Journal*, 62: 689–703.

MacMillan CP, Taylor L, Bi Y, Southerton SG, Evans R, Spokevicius A. 2015. The fasciclin-like arabinogalactan protein family of *Eucalyptus grandis* contains members that impact wood biology and biomechanics. *New Phytologist*, 206: 1314–1327.

Mahrez W, Shin J, Muñoz-Viana R, Figueiredo DD, Trejo-Arellano MS, Exner V, Siretskiy A, Gruissem W, Köhler C, Hennig L. 2016. BRR2a Affects Flowering Time via FLC Splicing. *PLoS Genetics*, 12: e1005924.

McKeown PC, Laouielle-Duprat S, Prins P, Wolff P, Schmid MW, Donoghue MTA, Fort A, Duszynska D, Comte A, Lao NT, Wennblom TJ, Smant G, Köhler C, Grossniklaus U, Spillane C. 2011. Identification of imprinted genes subject to parent-of-origin specific expression in *Arabidopsis thaliana* seeds. *BMC Plant Biology*, 11: 113.

McKown AD, Klápště J, Guy RD, Soolanayakanahally RY, La Mantia J, Porth I, Skyba O, Unda F, Douglas CJ, El-Kassaby YA, Hamelin RC, Mansfield SD, Cronk QCB. 2017. Sexual homomorphism in dioecious trees: extensive tests fail to detect sexual dimorphism in *Populus*. *Scientific Reports*, 7: 1831.

McManus CJ, Coolon JD, Duff MO, Eipper-Mains J, Graveley BR, Wittkopp PJ. 2010. Regulatory divergence in *Drosophila* revealed by mRNA-seq. *Genome Research*, 20: 816–825.

Meiklejohn CD, Coolon JD, Hartl DL, Wittkopp PJ. 2014. The roles of *cis*- and *trans*-regulation in the evolution of regulatory incompatibilities and sexually dimorphic gene expression. *Genome Research,* 24: 84–95.

Murase K Shigenobu S, Fujii S, Ueda K, Murata T, Sakamoto A, Wada Y, Yamaguchi K, Okasabe Y, Okasabe K, Kanno A, Ozaki Y, Takayama S. 2017. MYB transcription factor gene involved in sex determination in *Asparagus officinalis*. *Genes to Cells,* 22: 115–123.

Nagamitsu T, Futamura N. 2014. Sex expression and inbreeding depression in progeny derived from an extraordinary hermaphrodite of *Salix subfragilis*. *Botanical Studies,* 55: 1–9.

Pagnussat GC, Yu H-J, Ngo QA, Rajani S, Mayalagu S, Johnson CS, Capron A, Xie L-F, Ye D, Sundaresan V. 2005. Genetic and molecular identification of genes required for female gametophyte development and function in *Arabidopsis*. *Development,* 132(3): 603–614.

Parsch J, Ellegren H. 2013. The evolutionary causes and consequences of sex-biased gene expression. *Nature Review Genetics,* 14: 83–87.

Pucholt P, Rönnberg-Wästljung A-C, Berlin S. 2015. Single locus sex determination and female heterogamety in the basket willow (*Salix viminalis* L.). *Heredity,* 114: 575–583.

Pucholt P, Wright AE, Conze LL, Mank JE, Berlin S. 2017. Recent sex chromosome divergence despite ancient dioecy in the willow *Salix viminalis*. *Molecular Biology and Evolution,* 34: 1991–2001.

Puckett EE, Serapiglia MJ, DeLeon AM, Long S, Minocha R, Smart LB. 2012. Differential expression of genes encoding phosphate transporters contributes to arsenic tolerance and accumulation in shrub willow (*Salix* spp.). *Environmental and Experimental Botany,* 75: 248–257.

R Core Team. 2015. R: A language and environment for statistical computing. Vienna, Australia: R Foundation for Statistical Computing. ISBN: 3-900051-07-0. Available online at http://www.R-project.org/.

Rice, WR. 1992. Sexually antagonistic genes: experimental evolution. *Science*, 256: 1436–1439.

Rieseberg LH, Widmer A, Arntz AM, Burke JM. 2003. The genetic architecture necessary for transgressive segregation is common in both natural and domesticated populations. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 358: 1141–1147.

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26: 139–140.

Rodgers-Melnick E, Mane SP, Dharmawardhana P, Slavov GT, Crasta OR, Strauss SH, Brunner AM, DiFazio SP. 2012. Contrasting patterns of evolution following whole genome versus tandem duplication events in *Populus*. *Genome Research*, 22: 95–105.

Semerikov V, Lagercrantz U, Tsarouhas V, Rönnberg-Wästljung Alström-Rapaport Lascoux M. 2003. Genetic mapping of sex-linked markers in *Salix viminalis* L. *Heredity*, 91: 293–299.

Serapiglia MJ, Cameron KD, Stipanovic AJ, Smart LB. 2012. Correlations of expression of cell wall biosynthesis genes with variation in biomass composition in shrub willow (*Salix* spp.) biomass crops. *Tree Genetics and Genomes*, 8: 775–788.

Serapiglia MJ, Gouker FE, Hart JF, Unda F, Mansfield SD, Stipanovic AJ, Smart LB. 2014a. Ploidy level affects important biomass traits of novel shrub willow (*Salix*) hybrids. *BioEnergy Research*, 8: 259–269.

Serapiglia MJ, Gouker FE, Smart LB. 2014b. Early selection of novel triploid hybrids of shrub willow with improved biomass yield relative to diploids. *BMC Plant Biology*, 14: 74.

Shi X, Ng DWK, Zhang C, Comai L, Ye W, Chen J. 2012. *Cis*- and *trans*-regulatory divergence between progenitor species determines gene-expression novelty in *Arabidopsis* allopolyploids. *Nature Communications*, 3: 950.

Siriwardana CL, Gnesutta N, Kumimoto RW, Jones DS, Myers ZA, Mantovani R, Holt BF. 2016. NUCLEAR FACTOR Y, Subunit A (NF-YA) proteins positively regulate flowering and act through *FLOWERING LOCUS T*. *PLoS Genetics*, 12: e1006496.

Smart LB, Cameron KD. 2008. Genetic improvement of willow (*Salix* spp.) as a dedicated bioenergy crop. In: Vermerris W, editor. Genetic Improvement of Bioenergy Crops: Springer New York. pp. 377–396.

Song G, Guo Z, Liu Z, Cheng Q, Qu X, Chen R, Jiang D, Liu C, Wang W, Sun Y, Zhang L, Zhu Y, Yang D. 2013. Global RNA sequencing reveals that genotype-dependent allele-specific expression contributes to differential expression in rice F_1 hybrids. *BMC Plant Biology*, 13: 221.

Song Y, Ma K, Ci D, Chen Q, Tian J, Zhang D. 2013. Sexual dimorphic floral development in dioecious plants revealed by transcriptome, phytohormone, and DNA methylation analysis in *Populus tomentosa*. *Plant Molecular Biology*, 83: 559–576.

Springer NM, Stupar RM. 2007. Allele-specific expression patterns reveal biases and embryo-specific parent-of-origin effects in hybrid maize. *The Plant Cell*, 19: 2391–2402.

Stoof C, Richards BK, Woodbury PB, Fabio ES, Brumbach AR, Cherney J, Das S, Geohring L, Hansen J, Hornskey J, Mayton H, Mason C, Ruestow G, Smart LB, Volk TA, Steenhuis TS. 2014. Untapped potential: opportunities and challenges for sustainable bioenergy production from marginal lands in the Northeast USA. *BioEnergy Research*, 8: 482–501.

Stupar RM, Hermanson PJ, Springer NM. 2007. Nonadditive expression and parent-of-origin effects identified by microarray and allele-specific expression profiling of maize endosperm. *Plant Physiology*, 145: 411–425.

Stupar RM, Springer NM. 2006. *Cis*-transcriptional variation in maize inbred lines B73 and Mo17 leads to additive expression patterns in the F1 hybrid. *Genetics*, 173: 2199–2210.

Su X, Zeng B, Lin F, Qiao P, Qiaoli A, Huang W. 2016. How does long-term complete submergence influence sex ratio and resource allocation of a dioecious shrub, *Salix variegata* Franch.? *Ecological Engineering*, 87: 218–223.

Suvorov A, Nolte V, Pandey RV, Franssen SU, Futschik A, Schlötterer C. 2013. Intra-specific regulatory variation in *Drosophila pseudoobscura*. *PLoS ONE*, 8: e83547.

Taira T, Maëda J, Onishi T, Kitaura H, Yoshida S, Kato H, Ikeda M, Tamai K, Iguchi-Ariga SM, Ariga H. 1998. AMY-1, a novel C-MYC binding protein that stimulates transcription activity of C-MYC. *Genes to Cells*, 3: 549–565.

Tuskan GA, DiFazio SP, Faivre-Rampant P, Gaudet M, Harfouche A, Jorge V, Labbé JL, Ranjan P, Sabatti M, Slavov G, Street N, Tschaplinski TJ, Yin T. 2012. The obscure events contributing to the evolution of an incipient sex chromosome in *Populus*: a retrospective working hypothesis. *Tree Genetics and Genomes*, 8: 559–571.

Wang H, Mao Y, Yang J, He Y. 2015. *TCP24* modulates secondary cell wall thickening and anther endothecium development. *Frontiers in Plant Science*, 6: 436.

Wittkopp PJ, Haerum BK, Clark AG. 2004. Evolutionary changes in *cis* and *trans* gene regulation. *Nature*, 430: 85–88.

Wittkopp PJ, Haerum BK, Clark AG. 2008a. Regulatory changes underlying expression differences within and between *Drosophila* species. *Nature Genetics*, 40: 346–350.

Wittkopp PJ, Haerum BK, Clark AG. 2008b. Independent effects of *cis*- and *trans*-regulatory variation on gene expression in *Drosophila melanogaster*. *Genetics,* 178: 1831–1835.

Wittkopp PJ, Kalay G. 2011. *Cis*-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nature Review Genetics,* 13: 59–69.

Xu C, Bai Y, Lin Y, Zhao N, Hu L, Gong Z, Wendel JF, Liu B. 2014. Genome-wide disruption of gene expression in allopolyploids but not hybrids of rice subspecies. *Molecular Biology and Evolution*, 31: 1066–1076.

Yin T, DiFazio SP, Gunter LE, Zhang X, Sewell MM, Woolbright SA, Allan GJ, Kelleher CT, Douglas CJ, Wang M, Tuskan GA. 2008. Genome structure and emerging evidence of an incipient sex chromosome in *Populus*. *Genome Research*, 18: 422–430.

Zhai F, Mao J, Liu J, Peng X, Han L, Sun Z. 2016. Male and female subpopulations of *Salix viminalis* present high genetic diversity and high long-term migration rates between them. *Frontiers in Plant Science*, 7: 330.

Zhou R, Macaya-Sanz D, Rodgers-Melnick E, Carlson CH, Gouker FE, Evans LM, Schmutz J, Jenkins JW, Yan J, Tuskan GA, Smart LB, DiFazio SP. 2018. Characterization of a large sexually dimorphic genome interval in *Salix purpurea* L. (Salicaceae). *Molecular Genetics and Genomics,* DOI: 10.1007/s00438-018-1473-y.

CHAPTER 6

THE GENOMIC BASIS OF HETEROSIS IN HIGH-YIELDING TRIPLOID HYBRIDS OF SHRUB WILLOW (*SALIX* SPP.) BIOENERGY CROPS

6.1 Abstract

Many studies have highlighted the complex, multigenic basis for heterosis (hybrid vigor) in inbred crops. Despite the lack a consensus model, it is vital that we turn our attention to understanding heterosis in undomesticated, outcrossing, heterozygous, and often polyploid species, such as willow (*Salix* spp.). Shrub willow is a dedicated energy crop and is bred to be fast-growing and high-yielding on marginal land without competing with food crops. A trend in willow breeding is the consistent pattern of heterosis in triploid progeny produced from crosses between diploid and tetraploid species. Critical in understanding heterosis, the heritability of gene expression is dependent on allele-specific expression by local and remote factors in the genome. Here, I test whether differentially expressed genes are responsible for heterosis in triploid crosses made between diploid *S. purpurea*, diploid *S. viminalis*, and tetraploid *S. miyabeana* parents. Three biological replicates of progeny and parent shoot tips were collected after 11 weeks in the greenhouse and individually sequenced via RNA-Seq. My results highlight regulatory factors influencing differential expression and genes correlated with heterosis for traits collected in the greenhouse and in the field. Altogether, these data will be used to develop predictive models of heterosis and complement the growing genomic resources available for the improvement of shrub willow bioenergy crops.

6.2 Introduction

The regulation of gene expression has been attributed to both local *cis*-regulatory elements and distant *trans*-regulatory factors in the cell. Variation in these gene regulators can play dramatic roles in the evolution of gene expression. *Cis*-regulatory variation is thought to account for evolutionarily significant phenotypic differences, whereas *trans*-regulatory variation is thought to account more for adaptive differences (Wray, 2007). For instance, *cis*-regulatory variation in promoter regions within-species should be minimal, compared to that among species. So, it is more likely that *trans*-effects should account for most of the regulatory variation in the intraspecific hybrid, whereas *cis*-effects should account for most of the regulatory variation in the interspecific hybrid (Wittkopp et al., 2008; Wittkopp and Kalay 2012). More simply, the greater genetic distance between parents, the more likely it is that differential gene expression in the progeny will be due to gene localized polymorphism.

Many have focused on hybrids derived from crossing inbred parents (Guo et al., 2004; Guo et al., 2006), but few have focused on hybrids derived from outcrossing parents (Landry et al., 2005; Jhuang and Adams, 2007), and even fewer on hybrids derived from outcrossing parents of different species or ploidy (Wittkopp and Kalay 2012). From early studies based on only a few dozen genes to recent research employing Illumina RNA-Seq, a common result in maize, wheat, and rice is that there is a high frequency of additive gene expression in hybrids (Guo et al., 2006; Stupar and Springer 2006; Stupar et al., 2008; Wei et al., 2009), yet genes with nonadditive expression display allele-specific expression (Guo et al. 2004; Springer and Stupar 2007a; Wei et al., 2009). This differential expression could be due to remote *trans*-factors, whereby a small number of key regulatory genes can play significant roles in heterosis (Ni et al., 2009; He et al., 2010; Goff 2011).

In most crop plants, heterosis has been realized in hybrids bred from inbred parents of contrasting genetic backgrounds (East 1936; Birchler et al., 2003). For instance, in maize, high numbers of low-frequency alleles near conserved *cis*-regulatory regions in the genome have been thought to lead to gene misexpression and are implicated as having a deleterious impact on important component traits (Kremling et al., 2018). Dominance, exemplified by the complementation of deleterious parent alleles in the hybrid, may help explain the phenomenon of heterosis in maize, and there are efforts to purge these alleles from breeding material by implementing efficient targeted gene-editing technologies. Nevertheless, shrub willow are heterozygous outcrossers and often polyploid, so the genomic basis of heterosis is likely to be different from that of conventional inbred crop plants.

An Illumina-based reference genome assembly of female *S. purpurea* 94006 has been constructed using a F_2 map-guided approach to orient scaffolds into pseudomolecules (*Salix purpurea* v1.0, DOE-JGI, phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Spurpurea). The genome size is an estimated 400 Mb and contains approximately 37,000 primary gene models, and nearly twice the number of alternatively-spliced isoforms. Although there are numerous gaps in the *S. purpurea* v1 genome assembly, the reference has proven useful in read alignment, variant discovery, and candidate gene selection. There are a handful of studies in shrub willow focusing on genetic mapping (Gunter et al., 2003; Berlin et al., 2010; Hanley and Karp, 2016; Hallingback et al., 2016) of quantitative trait loci (QTL) associated with biomass-related traits to aid in marker-assisted selection (MAS), yet most have been low-resolution, utilizing fewer than 1000 markers.

Thus far, family-based allele-specific expression (ASE) in *Salix* is restricted to a single study of F_1 and F_2 intraspecific *S. purpurea* individuals (Carlson et al., 2017; see Chapter 5). For

both F_1 and F_2 families, expression-level dominance comprised the greatest nonadditive proportion of differentially expressed genes between parents. There was also an effect of sex on gene expression among F_1 individuals. A majority of those genes found to be differentially expressed localized to the sex determining region (SDR) on *S. purpurea* chr15. Overall, more genes with ASE were identified in the F_1 compared with the F_2, but both families displayed greater *cis*- than *trans*-regulatory divergent expression. This work aims to: 1) describe the inheritance and regulatory divergence patterns influencing gene expression within and among three interspecific F_1 triploid families, 2) test for dosage effects on parent alleles in hybrid progeny, and 3) determine whether nonadditive gene expression is correlated with biomass growth and wood chemical composition traits important for biomass production.

6.3 Materials and Methods

6.3.1 Plant Material and Growing Conditions

Progeny individuals from three full-sib F_1 triploid families included in this study were derived from the interspecific crosses: *S. purpurea* 94006 × *S. miyabeana* 01-200-003 (Family 415), *S. viminalis* 07-MGB-5027 × *S. miyabeana* 01-200-003 (Family 423), and *S. miyabeana* 01-200-006 × *S. viminalis* 'Jorr' (Family 430). Herein, we refer to parents of the F_1 families by their clone identifiers and discriminate the female and male parents as P1 and P2, respectively. Parent genotypes and randomly chosen progeny grown from stem cuttings (20 cm) in 12-L plastic pots with peat moss-based potting mix (Fafard, Agawam, MA) to evaluate growth traits under greenhouse conditions over the course of 12 weeks. Plot was defined as a single cutting planted in a pot, which were arranged in a randomized complete block design with four replicate blocks. Two blocks were located on benches in one greenhouse with the other two blocks in an adjacent greenhouse set for identical growing conditions. Supplemental greenhouse lighting was

provided on a 14-h day : 10-h night regimen with maximum daytime temperature of 26°C and a nighttime temperature of 18°C. Beyond weekly applications of beneficial insects and mites for pest management, no pesticides were required, as there were no symptoms of biotic or abiotic stress on any plant material throughout the length of the study. Liquid fertilizer (Peter's 15-16-17 Peat-Lite Special®, Scott's, Marysville, OH) was applied weekly after week four according to manufacturer recommendations.

6.3.2 Determination of Ploidy Level

The relative DNA content (pg $2C^{-1}$) of family parents and progeny was determined by flow cytometric analysis using young leaf material harvested from actively growing shoots in greenhouse conditions. Analysis of 50 mg of mature leaf tissue from parental genotypes and selected progeny was performed at the Flow Cytometry and Imaging Core Laboratory at Virginia Mason Research Center in Seattle, WA. A minimum of four replicates of all samples were independently assessed using the diploid female *S. purpurea* clone 94006 as an internal standard. Diploid parent clones from multiple runs were averaged and then divided by the value of the check for that run. This factor was then multiplied by each sample value within the same run as the check. When a clone was analyzed more than once, 2C values were averaged.

6.3.3 RNA sample preparation and sequencing

A total of three biological replicate shoot tips of all triploid progeny individuals, as well as their parents, were excised from the primary stem and immediately flash-frozen in liquid N_2 in the greenhouse, then placed in −80 °C storage. For each sample, a single shoot tip was removed from −80 °C storage, and ground to a fine powder (100-200 mg) prior to RNA isolation using the Spectrum™ Total Plant RNA Kit with DNase I digestion (Sigma, St. Louis, MO). Shoot tips were defined as the shoot axis that is the most distal part of a shoot system, comprised of a shoot

apical meristem and the youngest leaf primordia. The only modification to 'Protocol B' was that prior to the tissue lysis step, the 2-ME/lysate mixture was incubated at 65 °C for 5 min, otherwise, the manufacturers' procedures were followed. After elution, cold ethanol precipitations were performed by the addition of 10 µL acetic acid and 280 µL 100% cold ethanol to 100 µL eluate and placed in −80 °C for 3 h. Samples were centrifuged at 17,000 × *g* for 30 min at 4 °C, washed with 80% ethanol, then centrifuged at 17,000 × *g* for 20 min at 4 °C. After centrifugation, the supernatant was discarded and the pellet resuspended in ribonuclease-free 10 mM Tris-HCl. Quantification of RNA sample quality and concentration was performed using the Experion 'StdSens' kit (Bio-Rad Laboratories, Inc., Hercules, CA). Stranded RNA-Seq libraries were created and quantified by qPCR. Paired-end (2×76 bp or 2×151 bp) sequencing was performed on an Illumina Hi-Seq 2500 at J. Craig Venter Institute. Library sizes ranged from 8.3 to 53 million reads.

6.3.4 Read filtering, mapping, and variant discovery

Low-coverage paired-end genomic DNA sequencing of the parents of the F_1 families was performed to validate variants from RNA-Seq data. Biallelic SNPs were used to quantify allele-specific expression (ASE) within and among triploid progeny individuals. Parent DNA libraries were sequenced (Illumina HiSeq 2×101) and aligned to the *S. purpurea* v1 reference genome using BWA-MEM (Li and Durbin 2009). Subsequent SAM files were sorted, marked for duplicates, and indexed in Picard (broadinstitute.github.io/picard). Indel realignment and variant calling was performed using *HaplotypeCaller* (emit_conf=10, call_conf=30) in the Genome Analysis Toolkit (GATK) (DePristo et al., 2011).

Using *BBDuk* in the BBTools program (https://jgi.doe.gov/data-and-tools/bbtools/), raw reads were evaluated for artifact sequence by kmer matching (kmer = 25), allowing 1 mismatch

and detected artifact was trimmed from the 3' end of the reads. RNA spike-in reads, PhiX reads and reads containing any Ns were removed. Quality trimming was performed using the phred trimming method set at Q6. Finally, following trimming, reads under the length threshold were removed (≥ 25 bp or 1/3 original read length). BWA MEM (-M -p) was used for alignment of interleaved RNA-Seq reads to the *S. purpurea* reference. SAMtools was used to filter (-Shb -F 4 -f 0x2 -q 30), sort, and index resulting sequence alignment files. Duplicate reads were flagged using *MarkDuplicates* in Picard (broadinstitute.github.io/picard) and GATK was used to flag and realign indels with *RealignmentTargetCreator* (minReads = 20) and *IndelRealigner*.

6.3.5 Tests for Differential Expression

All statistical analyses were performed in the open-source statistical computing environment, R (R Core Team, 2015). Tests for differential expression were conducted in edgeR (Robinson et al., 2010). Normalization factors and dispersion estimates (robust = T) were computed prior to tests for differential expression using three biological replicate libraries of each genotype. A general linear model was used to fit normalized count data using *glmFit* and *glmLRT* to conduct likelihood ratio tests for the model coefficients. Tests for paired comparisons were conducted using an additive linear model using replicate within greenhouse as a fixed covariate. Genes were only considered to be differentially expressed at a false discovery rate (FDR) of 0.05.

6.3.6 Gene Expression Inheritance Classification

To determine the mode of gene expression inheritance, the number of RNA-Seq reads mapped to individual genes was counted for each of the female (P1) and male (P2) parents and progeny (H). Gene expression levels were compared based on normalized read counts in edgeR. Differentially expressed genes were determined using an Exact test for negative-binomially

distributed counts (FDR = 0.005). A custom R script was used to sort genes into the following six inheritance categories: P1-dominant (H≈P1 and H≠P2), P2-dominant (H≈P2 and H≠P1), additive (P1<H<P2 or P1>H>P2), overdominant (H>P1 and H>P2), underdominant (H<P1 and H<P2), and conserved (P1≈H≈P2).

6.3.7 Regulatory Divergence Classification

For regulatory divergence classifications, the sequence reads from progeny were assigned to each parent. For each gene, the expression levels of the two parental alleles were estimated based on allele counts across all SNP sites detected in parent DNA libraries (Carlson et al., 2017; see Chapter 5), allowing for the unambiguous assignment of parent origin. Categories of regulatory functions considered conserved, compensatory, ambiguous, *cis*, *trans*, *cis* + *trans*, and *cis* × *trans*, were assigned using R scripts, following previously described methods (Landry et al., 2005; McManus et al., 2010). Regulatory divergence assignments were based a binomial exact test between P1 and P2 alleles in the parents and between P1H and P2H alleles in the progeny, and a Fisher's Exact Test on P1, P2, P1H, and P2H.

6.3.8 Copy Number Variation

Copy number variation (CNV) was analyzed on a chromosome-wide scale, using median $\log_2$ (P2X / P4X) difference of logs in the parents and the median percentage of reads attributable to the diploid parent allele in the triploid hybrid. The expected copy number of each homeolog in the hybrid was either determined to be deficient, normal, or replete, depending on these two parameters, and is described in (Figure 6.1). To avoid over-estimating CNV in triploids, binned coverage of paired-end Illumina DNA-Seq reads of the parents was compared to validate RNA-Seq results.

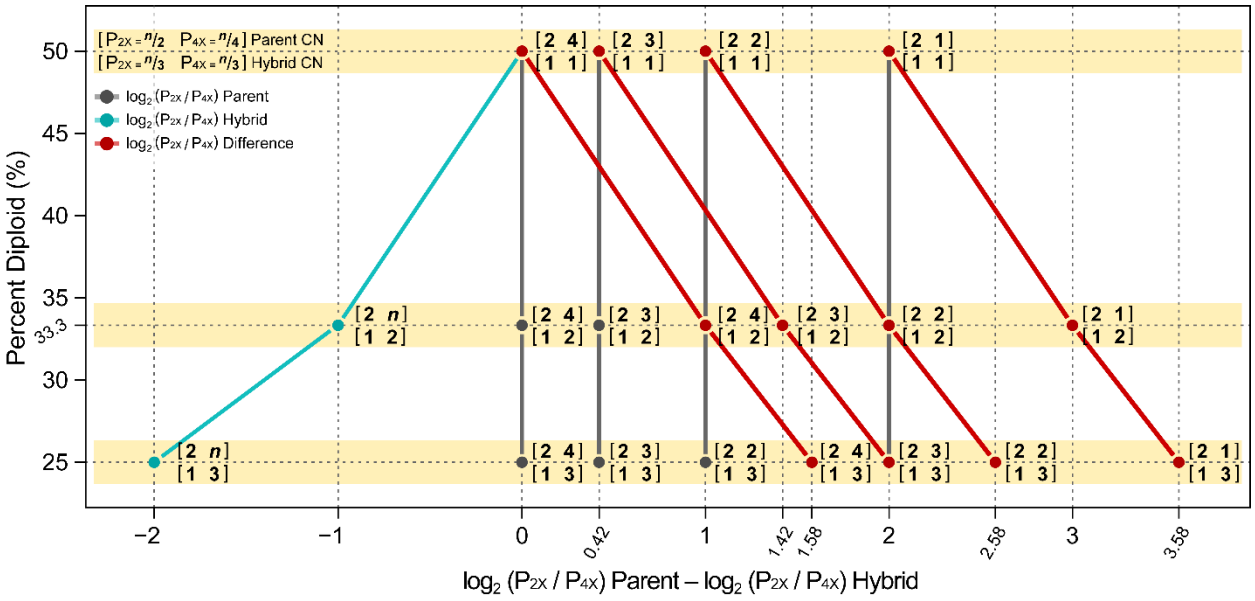

Figure 6.1 Schematic comparing parent allele-specific $\log_2$ (P2X / P4X) with the percentage of reads in a triploid hybrid attributable to the diploid parent (i.e., diploid %). Lines represent the $\log_2$ (P2X / P4X) in the hybrid (cyan), parent (dark grey), and the parent–hybrid difference (red). Brackets to the right of each point represents the copy number (CN) out of the total expected to occur in a diploid, triploid, or tetraploid state, e.g. [**2 4**] = [P_{2X} = **2**/2 *and* P_{4X} = **4**/4] in the parents and [**1 2**] = [P_{2X} = **1**/3 *and* P_{4X} = **2**/3] in the hybrid.

6.3.9 Gene Ontology Analysis

Gene ontology (GO)-term enrichment was performed in agriGO (Du et al., 2010) using the subset of the *S. purpurea* v1 transcriptome (reference set) that passed filtering, prior to tests of differential expression. Only significant ontologies (FDR = 0.05) were reported. *Salix purpurea* gene models and associated GO-terms which were annotated as hypothetical proteins were inferred using the best-hit (BLASTP E ≤ 0.1) to *Populus trichocarpa* (Phytozome v10.3) and *Arabidopsis thaliana* (TAIR10 and Araport11 annotations) proteomes.

6.3.10 Gene-Trait Correlations

Modules of highly-correlated genes as well as trait-module and intramodular correlations with hub genes were identified in the WGCNA package. Low-expressed genes (cpm < 1) were removed from expression tables prior to network construction, then normalized using the

varianceStabilizingTransformation function in DESeq2 (Love, et al. 2014). Co-expression modules were constructed as single blocks and assigned WGCNA color-codes. Gene-trait correlations were performed for each family using library-normalized expression values for genes and midparent heterosis for traits described in Chapter 3 (Table 3.2). Cumulative expression dysregulation was calculated according to Kremling et al. (2018).

6.4 Results

6.4.1 Transcriptome Analysis

After quality filtering and alignment of triploid F_1 progeny and parent paired-end RNA-Seq reads to the *S. purpurea* reference, the library sizes ranged from 10 to 56 million. Of the 105 libraries sequenced, there were two identified as outliers and removed prior to downstream statistical analyses: 13X-430-035 (greenhouse plot 59, biological replicate 1) and 12X-415-074 (greenhouse plot 278, biological replicate 3). To approximate the distances or leading $\log_2$ fold-changes between samples, library-normalized gene-level expression data was used in multi-dimensional scaling analysis on the top 500 common genes that represented the largest standard deviations between samples (Figure 6.2). The first MDS dimension represents sample distances based on species pedigree. For instance, individuals with *S. viminalis* in their background cluster to the left of the first dimension, *S. miyabeana* in the center, and *S. purpurea* to the right, such that *S. viminalis* parents (07-MBG-5027 and 'Jorr') and *S. purpurea* 94006 are at extremes, or the most distantly-related.

While family 415 and 423 individuals share the common tetraploid *S. miyabeana* parent 01-200-003, the proximity of family 423 and 430 clusters indicates that common parent species (*S. viminalis* and *S. miyabeana*) is a more important factor controlling transcriptome distances. The second dimension further separates samples by family. For all three triploid families, the

respective diploid and tetraploid parents flank clusters of family individuals, and are relatively equidistant from the offspring cluster centers. Taking both dimensions into account, Euclidean distances approximated here implies transcriptome-wide gene expression inheritance is mostly conserved or additive. However, this gene set only included those with the largest deviations common between all samples, whereas *pairwise* $\log_2$ fold-changes inform the extent of differential expression within-family.

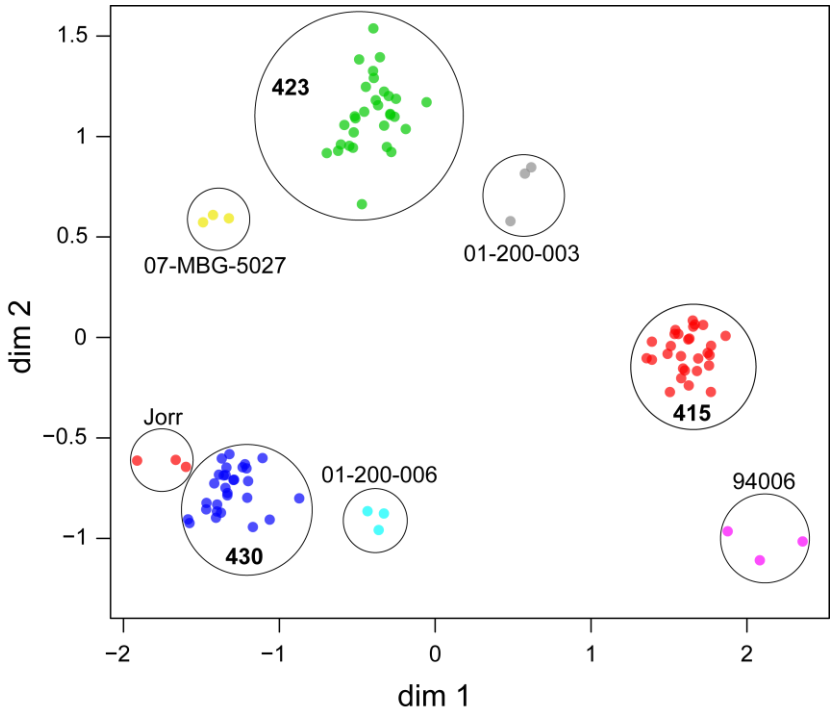

Figure 6.2 Multi-dimensional scaling (MDS) plot of library-normalized transcriptome-wide gene expression of all triploid F_1 progeny individuals (families 415, 423, and 430) and their diploid (94006, 07-MBG-5027, and 'Jorr') and tetraploid (01-200-006 and 01-200-003) parents. Euclidean distances on the two-dimensional plot approximate leading $\log_2$ fold-changes between samples, using the top 500 genes with the largest standard deviations.

6.4.2 Differential Gene Expression

Exact tests (FDR = 0.005) between triploid family parent genotypes yielded similar numbers of differentially expressed genes, but the P1:P2 ratios differed (Table 6.1). The comparison of family 415 parents, *S. purpurea* 94006 (P1) versus *S. miyabeana* 01-200-003

(P2), had 5166 differentially expressed genes, with 2,661 genes greater in 94006 and 2505 genes greater in 01-200-003 (P1:P2 = 1.06). The family 423 parents, *S. viminalis* 07-MBG-5027 (P1) versus *S. miyabeana* 01-200-003 (P2), had 5,523 differentially expressed genes, with 2,469 genes greater in 07-MBG-5027 and 3054 genes higher expressed in 01-200-003 (P1:P2 = 0.81). The family 430 parent comparison, *S. miyabeana* 01-200-006 (P1) versus *S. viminalis* 'Jorr' (P2), yielded 5,155 differentially expressed genes, with 2,467 genes greater in 01-200-006 and 2,688 genes greater in Jorr (P1:P2 = 0.91). Globally, the parents of family 423 had a greater percentage of genes that were differentially expressed (22.1%), compared to the parents of families 415 (20.8%) and 430 (20.5%).

Table 6.1 Number of differentially expressed genes between triploid family parents.

Family	Female (P1)	Male (P2)	P1 > P2 (%)	P1 < P2 (%)	P1 = P2 (%)	Total
415	94006	01-200-003	2,661 (10.7)	2,505 (10.1)	19,641 (79.2)	24,807
423	07-MBG-5027	01-200-003	2,469 (9.86)	3,054 (12.2)	19,519 (77.9)	25,042
430	01-200-006	Jorr	2,467 (9.81)	2,688 (10.7)	19,993 (79.5)	25,148

For those genes differentially expressed between parents, inheritance patterns were determined based on both the parent expression values and those observed in the hybrids (Table 6.2). For family 415, the percentage of differentially expressed genes showing nonadditive inheritance ranged from 27% to 39%, with an average of 33.5% (Figure 6.3). Percent nonadditive expression in family 423 ranged from 40% to 56%, with an average of 49.8%. Nonadditive expression in family 430 ranged from 34% to 60%, with an average of 50.3%.

Table 6.2 Number of genes assigned to inheritance classifications in triploid F_1 progeny individuals and their averages by family.

	P1-dominant	P2-dominant	Over-dominant	Under-dominant	Additive	Conserved
Family 415						
415-018	236	960	0	28	3,010	20,573
415-020	376	1,109	0	31	2,786	20,505
415-023	336	1,024	0	23	2,842	20,582
415-031	343	1,209	1	40	2,760	20,454
415-038	308	1,048	0	35	2,887	20,529
415-054	414	945	0	20	2,882	20,546
415-073	326	1,038	0	34	2,864	20,545
415-074	280	869	0	18	3,034	20,606
415-082	334	1,208	2	29	2,702	20,532
415-257	331	1,298	2	44	2,633	20,499
Mean	328.4	1,070.8	0.5	30.2	2,840	20,537.1
Family 423						
423-004	447	1,219	14	51	1,544	21,767
423-034	268	895	7	12	1,759	22,101
423-043	302	1,282	5	31	1,513	21,909
423-048	325	1,428	7	27	1,447	21,808
423-051	286	1,055	8	19	1,681	21,993
423-063	422	1,116	8	21	1,593	21,882
423-066	278	1,021	7	21	1,707	22,008
423-067	317	1,247	10	22	1,543	21,903
423-070	491	1,337	26	29	1,466	21,693
423-072	324	1,332	3	34	1,495	21,854
Mean	346	1,193.2	9.5	26.7	1,574.8	21,891.8
Family 430						
430-004	1,498	370	5	41	1,330	21,904
430-005	1,517	340	4	29	1,249	22,009
430-006	867	274	5	22	1,609	22,371
430-016	1,067	285	7	16	1,492	22,281
430-018	1,162	291	6	14	1,410	22,265
430-025	1,175	283	2	25	1,380	22,283
430-031	1,026	279	6	20	1,482	22,335
430-033	1,007	458	6	31	1,496	22,150
430-034	1,519	351	14	21	1,295	21,948
430-035	704	207	3	16	1,780	22,438
Mean	1,154.2	313.8	5.8	23.5	1,452.3	22,198.4

Transgressively expressed genes (overdominant or overdominant) averaged just 0.7%, 1.1%, and 1%, for families 415, 423, and 430, respectively. The percentage of genes with underdominant expression out of total transgressively expressed genes was 98%, 74%, and 80% for families 415, 430, and 415, respectively. All individuals had a higher percentage of genes with dominant expression in the direction of the tetraploid parent, ranging from 66% to 88% and averaging 70% across all triploid families.

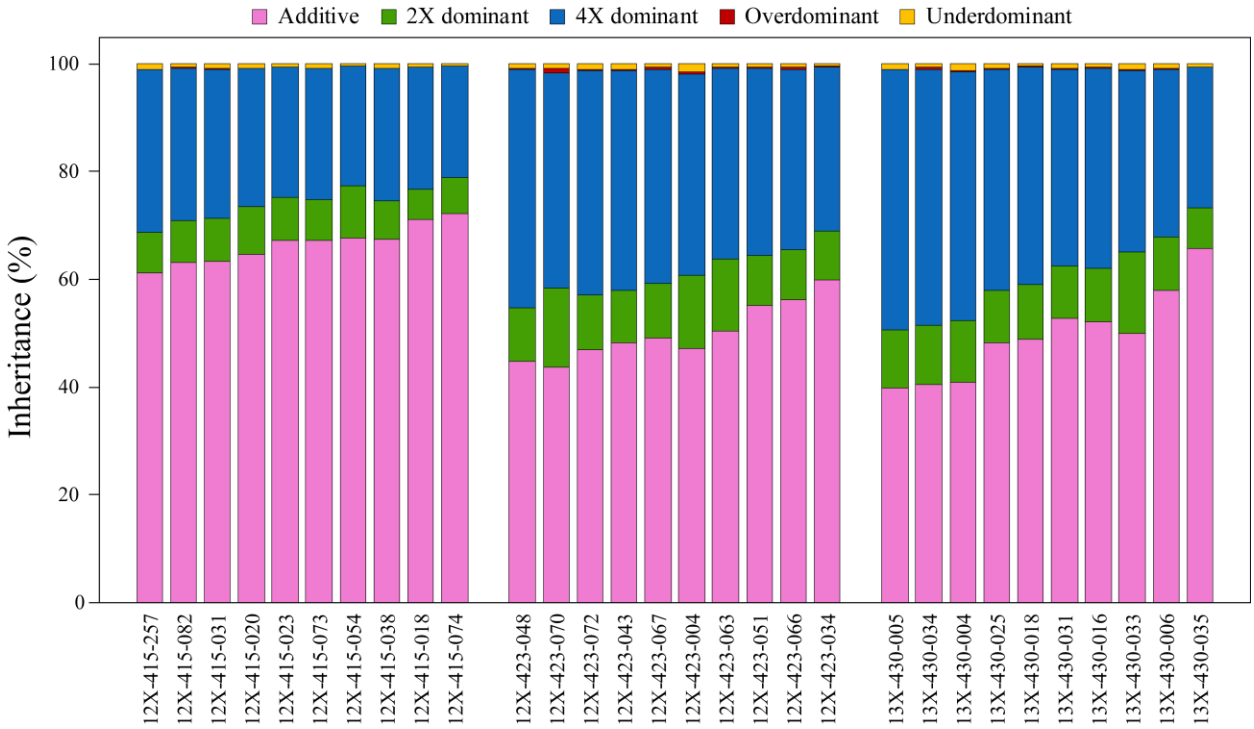

Figure 6.3 Percentage of additive and nonadditive genes in triploid individuals. Only genes differentially expressed between the parents (not conserved or ambiguous) were used in to calculate inheritance class percentages. Gene expression inheritance patterns are colored according to the legend to the right of the graph. Individuals are shown in order of their additive (%) of differentially expressed genes.

There were fewer numbers of diploid parent dominant genes (15) (Table 6.3) than tetraploid parent dominant genes (89) (Table 6.4) that were common across all families and individuals. Due to the low number of common diploid parent dominant genes, there were no significant functional enrichments. Tetraploid dominant genes were enriched for GO molecular functions: beta-glucosidase activity (GO:0008422) and catalytic activity (GO:0003824) (Table 6.4). In addition, tetraploid parent dominant genes were enriched for the KEGG pathways: phenylpropanoid biosynthesis (K00940), cyanoamino acid metabolism (K00460), biosynthesis of secondary metabolites (K01110), metabolic pathways (K01100), and starch and sucrose metabolism (K00500) (Table 6.5).

Table 6.3 Common diploid parent dominant genes across all triploid family individuals.

Gene	Chr	Mb	Gene Annotation	log_2CPM
SapurV1A.1417s0070	6	10.261	hypothetical protein	2.6
SapurV1A.1417s0030	6	10.308	hypothetical protein	1.5
SapurV1A.0659s0110	7	4.584	brassinosteroid insensitive 1	1.2
SapurV1A.0557s0080	10	14.137	UDP-glucosyltransferase 91A1	2.1
SapurV1A.0243s0410	10	17.997	Sec12-like protein 1	1.8
SapurV1A.0026s0110	11	8.065	ankyrin repeat protein	2.5
SapurV1A.0026s0120	11	8.100	ankyrin repeat protein	3.0
SapurV1A.0382s0200	14	4.163	alanine-tRNA ligase	1.2
SapurV1A.0052s0180	14	9.675	COBRA-like protein 7	1.7
SapurV1A.0154s0300	16	0.254	UDP-glucose flavonoid 3-O-glucosyltransferase	4.6
SapurV1A.0431s0060	16	9.053	isoamyl acetate-hydrolyzing esterase	1.8
SapurV1A.1116s0110	17	17.951	disease resistance RPP13-like protein 1	4.4
SapurV1A.0654s0010	654	0.018	UDP-glucosyltransferase 85A2	5.0
SapurV1A.1047s0020	1047	0.023	DEAD-box ATP-dependent RNA helicase 37	0.7
SapurV1A.4767s0010	4767	0.006	hypothetical protein	3.8

Table 6.4 Common tetraploid parent dominant genes across all triploid family individuals.

Gene	Chr/Scaff	Mb	Gene Annotation	log_2 CPM
SapurV1A.0280s0010	1	2.776	beta-amyrin synthase	4.2
SapurV1A.0094s0530	1	5.302	hypothetical protein	0.7
SapurV1A.0267s0350	2	4.860	2-oxoisovalerate dehydrogenase alpha subunit	2.4
SapurV1A.0155s0140	2	18.308	Phi class glutathione transferase GSTF5	6.4
SapurV1A.1849s0020	3	0.028	hypothetical protein	3.3
SapurV1A.0059s0740	3	12.728	hypothetical protein	4.5
SapurV1A.0009s1090	3	15.313	hypothetical protein	2.5
SapurV1A.0028s0020	4	1.803	cotton fiber	2.3
SapurV1A.1124s0060	5	1.538	tau class glutathione transferase GSTU58	5.4
SapurV1A.1571s0080	5	2.789	RING/U-box superfamily protein MBR1	3.4
SapurV1A.0539s0260	5	18.201	chlorophyllase	2.1
SapurV1A.0001s2410	6	5.866	oxidosqualene cyclase	2.7
SapurV1A.0001s1650	6	6.407	dihydroflavonol-4-reductase	2.7
SapurV1A.0341s0030	6	15.772	GDSL-lipase	8.1
SapurV1A.0710s0050	6	21.582	transparent testa glabra2	5.3
SapurV1A.0176s0500	8	6.800	floral homeotic protein AP1	2.9
SapurV1A.0580s0110	8	11.446	hAT family dimerization domain protein	1.5
SapurV1A.0206s0440	8	13.107	sieve element occlusion protein	1.7
SapurV1A.0330s0190	10	4.939	remorin, carboxy-terminal region protein	4.1
SapurV1A.0330s0260	10	5.003	HXXXD-type acyl-transferase-like protein	4.5
SapurV1A.0330s0270	10	5.008	HXXXD-type acyl-transferase-like protein	3.3
SapurV1A.1418s0080	10	5.013	HXXXD-type acyl-transferase-like protein	2.7
SapurV1A.0518s0140	10	17.249	pleckstrin-like (PH) and lipid-binding START domain	2.1
SapurV1A.0036s0120	10	18.313	oxalate oxidase, putative	8.3
SapurV1A.0036s0140	10	18.328	RmlC-like cupins superfamily protein	11.1
SapurV1A.1533s0080	11	0.759	MATE efflux family protein	10.6
SapurV1A.1533s0050	11	0.770	adenosine deaminase	5.9
SapurV1A.0679s0080	11	12.373	ankyrin repeat protein	3.9
SapurV1A.0527s0200	11	17.285	reticuline oxidase	4.9
SapurV1A.0202s0140	12	2.133	reticuline oxidase	1.9
SapurV1A.1070s0030	12	13.417	GDSL-lipase	6.6
SapurV1A.1070s0020	12	13.424	GDSL-lipase	6.4
SapurV1A.1196s0110	13	4.832	L-tyrosine decarboxylase	5.4
SapurV1A.0182s0060	13	9.083	target of Myb protein 1	5.8

Table 6.4
(continued)

Gene	Chr/Scaff	Mb	Gene Annotation	log_2 CPM
SapurV1A.1866s0020	13	12.615	isoflavone reductase-P3	4.6
SapurV1A.0383s0160	13	12.990	isoflavone reductase-P3	6.9
SapurV1A.1156s0070	14	3.752	cytochrome P450 family protein	2.4
SapurV1A.0498s0210	15	2.070	HAT family dimerization protein	3.6
SapurV1A.0354s0280	15	13.271	myb-like DNA-binding domain protein	2.5
SapurV1A.0011s0470	15	17.984	tryptophan aminotransferase	4.8
SapurV1A.1225s0140	15	22.518	MATE efflux family protein	5.9
SapurV1A.0224s0010	16	5.047	glycoside hydrolase family 1 protein	1.5
SapurV1A.0008s0010	16	11.863	beta-glucosidase	2.9
SapurV1A.0008s0190	16	12.020	beta-glucosidase	4.9
SapurV1A.0005s0350	16	17.270	porin/voltage-dependent anion-selective channel protein	2.6
SapurV1A.1040s0100	16	25.150	receptor-like Serine/Threonine kinase	3.6
SapurV1A.0064s0010	16	25.265	polyphenol oxidase	6.5
SapurV1A.0064s0050	16	25.347	polyphenol oxidase	2.4
SapurV1A.0312s0020	17	2.675	RING-H2 zinc finger protein	0.9
SapurV1A.1999s0070	17	2.981	UDP-glucose flavonoid 3-O-glucosyltransferase	12.3
SapurV1A.0871s0040	17	5.145	F-box protein interaction domain protein	2.9
SapurV1A.0230s0160	17	15.348	phosphoribosylformimino-5-aminoimidazole carboxamide isomerase	7.0
SapurV1A.0314s0200	18	0.934	aspartyl protease	6.1
SapurV1A.0435s0070	18	5.303	Myb-like transcription factor glabra2	4.9
SapurV1A.0581s0030	18	8.135	cytosolic aldehyde dehydrogenase	5.7
SapurV1A.0724s0160	18	9.319	GDSL esterase/lipase	9.1
SapurV1A.0940s0020	18	9.414	GDSL esterase/lipase	7.6
SapurV1A.0034s0420	18	13.811	plastocyanin-like domain protein	0.9
SapurV1A.0553s0020	19	2.805	NBS-LRR resistance protein	2.8
SapurV1A.1054s0010	19	3.331	glycoside hydrolase family 1 protein	1.3
SapurV1A.0305s0160	305	0.203	ankyrin repeat protein	2.8
SapurV1A.0313s0200	313	0.271	GDSL-like lipase/acylhydrolase	8.5
SapurV1A.0395s0260	395	0.231	MATE efflux family protein	4.6
SapurV1A.0427s0090	427	0.092	transposase-associated domain protein	2.9
SapurV1A.0519s0030	519	0.023	beta-glucosidase	5.8
SapurV1A.0537s0240	537	0.166	flavonol 4'-sulfotransferase	2.9
SapurV1A.0537s0260	537	0.190	flavonol 4'-sulfotransferase	3.4

Table 6.4
(continued)

Gene	Chr/Scaff	Mb	Gene Annotation	log_2 CPM
SapurV1A.0733s0010	733	0.010	lactoylglutathione lyase-like lyase	5.4
SapurV1A.0799s0040	799	0.043	ankyrin repeat protein	3.9
SapurV1A.0799s0060	799	0.067	ankyrin repeat protein	5.1
SapurV1A.0830s0010	830	0.064	nuclear-interacting partner of ALK	5.5
SapurV1A.0856s0020	856	0.027	hypothetical protein	1.7
SapurV1A.0941s0030	941	0.049	jacalin-like lectin domain protein	9.7
SapurV1A.0954s0060	954	0.075	NBS-LRR resistance protein	1.9
SapurV1A.1020s0050	1020	0.087	Serine/Threonine-kinase SAPK10	2.3
SapurV1A.1269s0010	1269	0.016	polarity axis stabilization protein, Afi1 N-terminal	2.2
SapurV1A.1551s0030	1551	0.029	transmembrane protein	6.3
SapurV1A.1572s0030	1572	0.037	Serine/Threonine kinase domain protein	4.3
SapurV1A.1612s0020	1612	0.026	CC-NBS-LRR resistance protein	1.8
SapurV1A.1631s0010	1631	0.013	serine carboxypeptidase	3.5
SapurV1A.1723s0100	1723	0.045	dirigent-like protein pDIR11	2.4
SapurV1A.1772s0030	1772	0.034	S-adenosyl-L-methionine:salicylic acid carboxyl methyltransferase	4.1
SapurV1A.1878s0010	1878	0.011	plastocyanin-like domain protein	0.6
SapurV1A.2455s0030	2455	0.018	TIR-NBS-LRR disease resistance-like protein	2.7
SapurV1A.2466s0010	2466	0.008	lipoxygenase	9.5
SapurV1A.3023s0040	3023	0.012	P-coumaroyl shikimate 3'-hydroxylase	6.1
SapurV1A.3977s0030	3977	0.009	transmembrane protein	3.5
SapurV1A.4321s0010	4321	0.003	UDP-glucosyltransferase	3.8
SapurV1A.4344s0010	4344	0.003	phytol kinase	4.4

Table 6.5 Functional enrichments for tetraploid-parent dominant genes.

Pathway ID	Pathway Description	Count	FDR
GO Molecular Function			
GO:0008422	beta-glucosidase activity	4	0.0339
GO:0003824	catalytic activity	28	0.0448
KEGG Pathways			
K00940	phenylpropanoid biosynthesis	6	0.0002
K00460	cyanoamino acid metabolism	4	0.0007
K01110	biosynthesis of secondary metabolites	10	0.0041
K01100	metabolic pathways	13	0.0097
K00500	starch and sucrose metabolism	4	0.0269

6.4.3 Allele-Specific Expression

In order to determine the extent of regulatory divergent expression in the three triploid families, tests for ASE were conducted using expression data on biallelic sites that were first called with parent DNA-Seq and RNA-Seq libraries prior to calling parent alleles in the progeny. Family averages for the total number of genes assigned to at least one regulatory class were 15,391 (±114), 16,800 (±72), and 16,711 (±113), for families 415, 423, and 430, respectively (Table 6.6). On average, the percentage of genes assigned to non-conserved regulatory classes was 12%, 11%, and 10%, for families 415, 423, ad 430, respectively. Family 415 had the greatest percentage of non-conserved genes with *cis*-regulation (65%), compared to families 423 (58%) and 430 (54%). The greatest mean percentage of genes with *trans* (24.6%), *cis* × *trans* (7.4%), and compensatory (10.8%) regulatory divergence patterns was for family 430, whereas family 415 had the greatest mean *cis* + *trans* (5.1%). Across all triploid individuals, a total of 49 genes were in common, having either *cis*, *trans*, *cis* + *trans*, *cis* × *trans*, or compensatory regulatory classifications (Table 6.7).

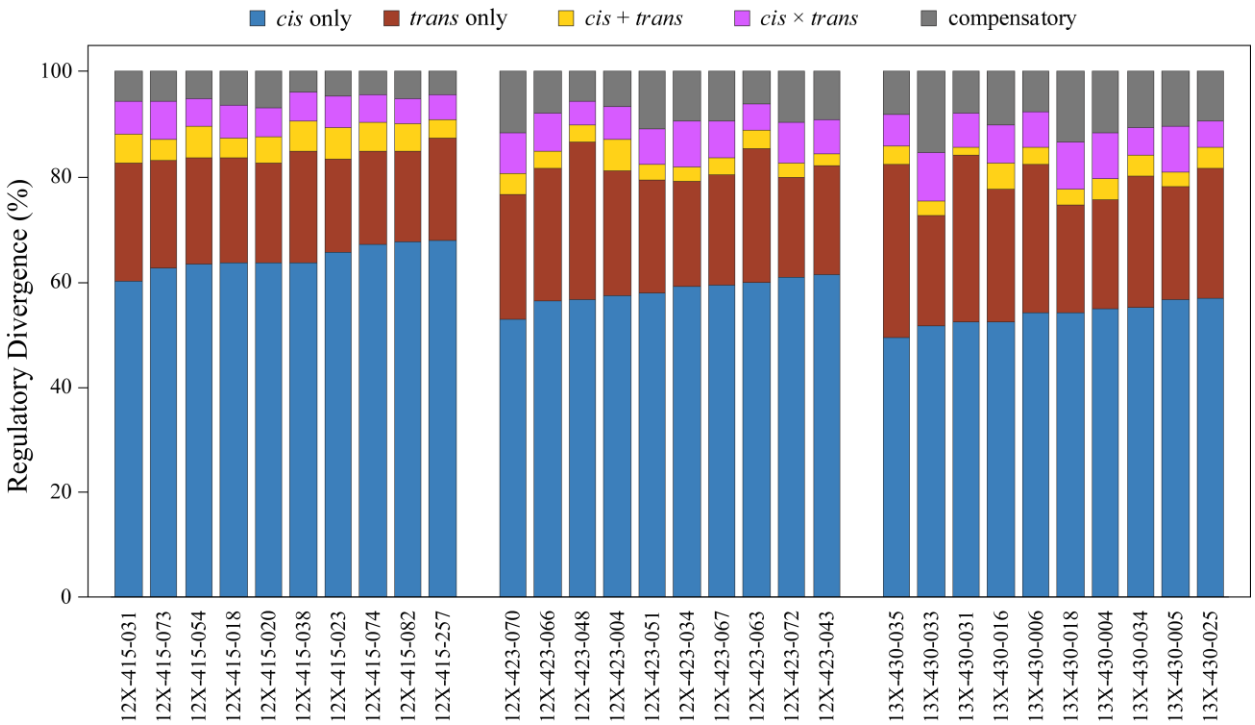

Figure 6.4 Regulatory divergence percentages by class. Only genes showing allele-specific expression (not conserved or ambiguous) are depicted. Classes are colored according to the legend above the graph. Individuals within families are ordered according to cis (%).

In addition, higher proportions of overdominant and underdominant expression coincided with higher proportions of *cis* × *trans* and compensatory regulatory classes. Further, a higher proportion of *cis* + *trans* divergence coincided with a lower proportion of underdominant expression, most notably for the comparison of families 423 and 430 with family 415.

Table 6.6 Number of genes assigned to regulatory divergence classifications (FDR = 0.005) in triploid F_1 individuals and their means by family.

	cis	***trans***	***cis* + *trans***	***cis* × *trans***	**Compensatory**	**Ambiguous**	**Conserved**
Family 415							
415-018	152	210	21	34	36	1,151	13,692
415-020	412	123	32	36	44	1,084	13,447
415-023	384	105	34	35	28	1,084	13,984
415-031	316	119	29	32	30	1,109	14,004
415-038	328	109	29	29	20	1,089	13,802
415-054	381	122	35	32	31	1,049	13,839
415-073	354	116	23	41	32	1,097	14,025
415-074	288	76	24	22	19	1,201	12,841
415-082	359	91	27	25	28	1,125	13,924
415-257	329	95	17	22	22	1,131	13,896
Mean	330	117	27	31	29	1,112	13745
Family 423							
423-004	252	105	26	28	29	1,224	15,071
423-034	317	108	14	48	50	1,219	15,116
423-043	306	103	11	33	45	1,248	15,112
423-048	178	94	10	14	18	1,284	15,472
423-051	316	117	16	38	59	1,225	15,051
423-063	292	124	17	24	30	1,218	15,096
423-066	187	84	11	24	26	1,351	14,519
423-067	283	99	16	33	45	1,238	15,094
423-070	252	113	20	36	56	1,237	15,199
423-072	274	85	13	35	43	1,261	15,216
Mean	266	103	15	31	40	1,251	15,095
Family 430							
430-004	295	111	21	47	63	1,095	14,797
430-005	203	77	10	32	37	1,110	15,386
430-006	219	115	13	27	31	1,095	15,298
430-016	245	117	24	34	47	1,086	15,501
430-018	252	95	14	42	62	1,118	15,210
430-025	180	79	12	16	30	1,182	15,615
430-031	174	105	5	22	26	1,171	15,366
430-033	265	107	14	47	79	1,126	14,884
430-034	206	94	15	19	40	1,126	15,301
430-035	144	96	10	17	24	1,187	14,399
Mean	218	100	14	30	44	1,130	15,176

Table 6.7. Genes with *cis*, *trans*, or *cis–trans* compensatory regulatory classifications (i.e., not conserved or ambiguous), across all 3 triploid families. Unique regulatory classes observed for each gene within a family are separated by a comma.

Gene	Chr	Start	Gene Annotation	logCPM	Family 415	Family 423	Family 430
SapurV1A.0820s0070	1	3.257	naringenin-chalcone synthase	9.4	*trans*	*cis* + *trans*	*cis* + *trans*
SapurV1A.0260s0310	1	7.746	anthocyanidin synthase	8.5	*trans*	*cis* + *trans*	*cis* + *trans*
SapurV1A.1365s0070	1	17.426	light harvesting chlorophyll A-B-binding	10.1	*trans*	*cis* × *trans*	compensatory
SapurV1A.0004s0940	2	8.526	auxin-responsive protein (Aux/IAA)	9.1	*cis*	*cis*	*cis*
SapurV1A.0014s0660	2	19.863	PSI light harvesting chlorophyll A/B-binding	11.4	*cis* + *trans*	compensatory	*cis*
SapurV1A.0937s0100	2	22.463	hypothetical protein	9.1	*cis* + *trans*	*cis* × *trans*, *trans*	*cis*
SapurV1A.0937s0110	2	22.47	transmembrane protein	9.3	*trans*	*cis*	*cis*
SapurV1A.0056s0640	3	14.482	naringenin-chalcone synthase	7.9	*cis*	*cis* + *trans*	*cis* × *trans*
SapurV1A.0610s0050	4	0.582	MYB transcription factor	4.7	*cis*	*trans*	*trans*
SapurV1A.0722s0220	4	0.984	thiamine thiazole synthase	8.9	*trans*	*cis* × *trans*	*cis* × *trans*
SapurV1A.0315s0080	4	6.392	ribulose bisphosphate carboxylase small chain	11.7	*trans*	*cis* + *trans*	*trans*
SapurV1A.1318s0090	4	9.608	Non-specific lipid-transfer protein	9.7	*cis*	*cis*	*cis*
SapurV1A.0065s0400	4	18.062	Non-specific lipid-transfer protein	7.4	*cis*	*cis* × *trans*	*cis*
SapurV1A.1038s0020	5	0.186	plant/F25P12-18 protein	10.0	*cis*	*cis* × *trans*	*cis*
SapurV1A.0266s0210	5	22.609	glyceraldehyde-3-phosphate dehydrogenase	9.9	*trans*	compensatory	*cis* × *trans*
SapurV1A.0359s0030	6	0.767	phloem protein 2-A10	5.9	*cis*	*cis*	compensatory
SapurV1A.0141s0150	6	12.004	phenylalanine ammonia-lyase	10.5	*cis*	*cis* + *trans*	*cis*
SapurV1A.0616s0140	6	13.171	auxin-binding protein ABP19a	11.2	*trans*	*cis* × *trans*	*cis* × *trans*
SapurV1A.0166s0340	7	1.174	histone H4	8.8	*cis*	*cis* × *trans*	*cis*
SapurV1A.0415s0130	9	4.045	class I small heat shock protein	6.1	*cis*	*cis*	compensatory
SapurV1A.0027s0890	9	4.645	protein transporter SEC31	7.3	*cis*	*cis*	*cis*
SapurV1A.0042s0020	9	6.225	CHCH domain protein	8.0	*cis*	*cis*	*cis*
SapurV1A.2136s0020	9	7.492	Serine/Threonine kinase, plant-type protein	8.1	*cis* + *trans*	*trans*	*trans*
SapurV1A.1150s0050	9	10.474	aquaporin, MIP family, PIP subfamily	8.4	*cis*, *trans*	*cis*	*trans*
SapurV1A.0072s0630	10	11.825	stress responsive A/B barrel domain	7.1	*cis*	*cis*, *cis* × *trans*	*cis*
SapurV1A.0023s0970	10	16.458	glutathione S-transferase DHAR2	8.8	compensatory	*cis*	*trans*
SapurV1A.0023s0850	10	16.54	chalcone-flavonone isomerase 1	8.5	*cis* + *trans*	compensatory	compensatory
SapurV1A.0243s0490	10	17.936	histone H2B	8.9	*cis*	*cis*	*cis*
SapurV1A.0047s0140	11	15.535	tau class glutathione transferase GSTU8	8.3	*cis*, *cis* + *trans*	*cis*	*cis*
SapurV1A.0989s0030	12	3.892	pathogenesis-related protein bet V I family	8.0	*trans*	*cis* × *trans*	*cis*
SapurV1A.1070s0180	12	13.339	60S ribosomal protein L24	6.6	*cis*	*cis*	*cis*
SapurV1A.0920s0090	13	0.288	transmembrane protein	4.5	*cis*	*cis*	compensatory

Table 6.7
(continued)

Gene	Chr	Start	Gene Annotation	logCPM	Family 415	Family 423	Family 430
SapurV1A.0088s0190	13	0.491	plant/F25P12-18 protein	9.5	*cis* + *trans*	*cis*	*cis* × *trans*
SapurV1A.0092s0020	13	2.383	cytochrome P450 family flavone synthase	9.3	*cis*	*cis*	*cis*
SapurV1A.0706s0160	14	1.043	progesterone 5 beta-reductase	9.0	*cis*	*cis*	*cis*
SapurV1A.0142s0220	14	12.239	GDSL esterase/lipase	8.8	*trans*	*cis* + *trans*	*trans*
SapurV1A.0322s0060	17	12.327	latex abundant protein	9.4	*cis* + *trans*	*cis*	*cis*
SapurV1A.0101s0240	17	14.514	ribulose bisphosphate carboxylase small chain	12.3	*cis* + *trans*	*trans*	*trans*
SapurV1A.0192s0200	18	2.238	protease inhibitor/seed storage/LTP family	9.2	*cis* + *trans*	*cis* × *trans*, *trans*	*trans*
SapurV1A.0034s0220	18	13.621	protease inhibitor/seed storage/LTP family	9.2	*trans*	*cis* + *trans*	*cis*, *cis* + *trans*
SapurV1A.0034s0800	18	14.077	Grx-C4-glutaredoxin	7.6	*cis*	*cis*	*cis*
SapurV1A.0054s0230	54	0.212	PSI light harvesting chlorophyll A/B-binding	7.3	*cis*	*cis*	*cis*
SapurV1A.0502s0110	502	0.11	oxygen-evolving enhancer protein	9.0	*cis*	*cis*	*cis* × *trans*
SapurV1A.0757s0060	757	0.043	plasma membrane intrinsic protein	9.5	*trans*	*cis*	*cis* × *trans*
SapurV1A.1208s0070	1208	0.044	UDP-glycosyltransferase	8.3	*cis*	*cis*	*cis*
SapurV1A.1305s0010	1305	0.002	kunitz trypsin inhibitor TI3	8.6	*cis*	*cis* × *trans*	*trans*
SapurV1A.2349s0020	2349	0.009	protease inhibitor/seed storage/LTP family	7.6	*cis* × *trans*	*cis*	*cis*
SapurV1A.3012s0030	3012	0.013	naringenin-chalcone synthase	9.1	*cis* + *trans*	*cis*	*trans*
SapurV1A.3261s0010	3261	0.007	microsomal glutathione S-transferase	6.7	*cis*	*cis*	compensatory

6.4.4 Sex-Biased Gene Expression

The number of genes differentially expressed (FDR = 0.005) by sex were 181 and 83 for families 415 and 423, respectively. Comparisons could not be made for family 430 and was not included in the analysis because all individuals were identified as female. Approximately equal numbers of genes displayed greater expression in either sex (females = 90, males = 91) in family 415, but in family 423, females had greater numbers of genes with greater expression (females = 51, males = 32). A subset of these genes are listed in Table 6.8. Many of these genes are located within the non-recombining, sex-determining region on chr15, or scaffolds which represent Z or W alternative haplotypes of chr15 (Figure 6.5). While there were no overrepresented GO-terms in either male- or female-biased gene sets, there is evidence many genes within these sets interact and could be a part of larger networks (Figure 6.6). The female gene set included a number of flowering pathway genes, including the floral homeotic genes *APETALA 1* and *PISTILLATA*, whereas kinesin-like, kinases, and those genes involved in secondary metabolism were mostly found in the male set. Genes differentially expressed by sex identified in these triploid families were also recognized within the F_1 intraspecific *S. purpurea* family 82 (Carlson et al., 2017; see Chapter 5). The most prominent of these genes were: activating signal cointegrator 1 complex subunit 3 (SapurV1A.2212s0020 and SapurV1A.2212s0030), stomatal cytokinesis defective 1 (SapurV1A.4349s0010 and SapurV1A.1896s0010), C3HC zinc finger-like nuclear interacting partner of ALK (SapurV1A.0582s0010), pre-mRNA-splicing factor ATP-dependent RNA helicase DEAH1 (SapurV1A.1479s0010), poly(A) polymerase 1 nucleotidyltransferase (SapurV1A.0475s0010), Wuschel homeobox 1 (SapurV1A.0718s0010), and five encoding kinesin-like proteins (SapurV1A.4040s0010, SapurV1A.1267s0020, SapurV1A.1267s0010, SapurV1A.0719s0090, SapurV1A.1267s0030).

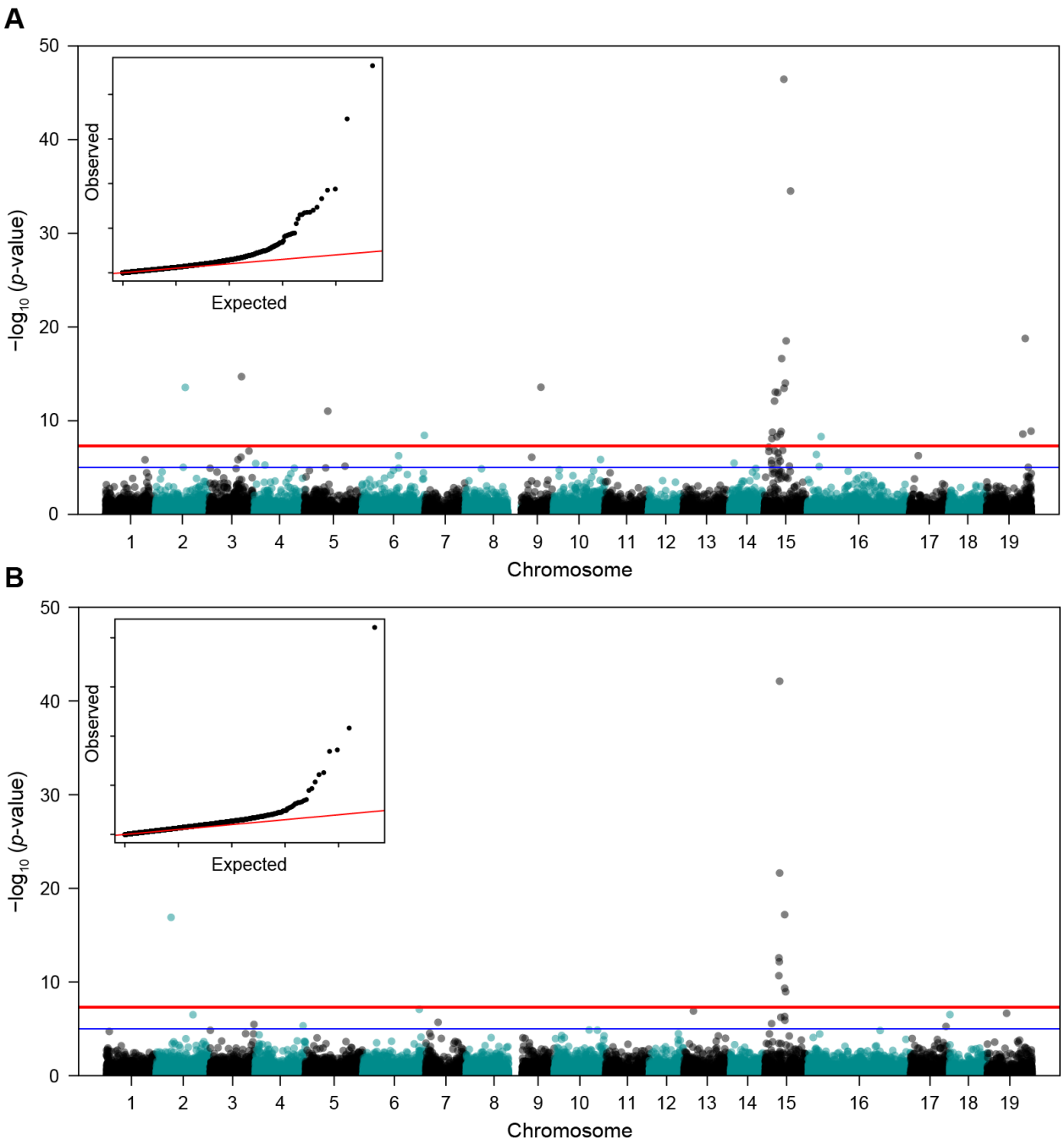

Figure 6.5 Manhattan plots of differentially expressed genes between three female and three male individuals in each of **(A)** family 415 (*S. purpurea* × *S. miyabeana*) and **(B)** family 423 (*S. viminalis* × *S. miyabeana*). Quantile-quantile (QQ)-plots in the top left of each panel assess model fit and depicts observed and expected $-\log_{10}$ (*p*-values). The blue and red horizontal lines in both panels represent genome-wise significance at a $-\log_{10}$ (1×10^{-5}) and $-\log_{10}$ (1×10^{-8}), respectively.

Table 6.8 Genes differentially expressed by sex ($|\log_2 FC| > 1.2$, $-\log_{10}(p\text{-value}) > 4$) in families 415 and 423. Family 430 was not included, as all individuals in the family were female. Unplaced *S. purpurea* v1 scaffolds are those > 19.

***Salix purpurea* v1 Gene Model**	**Chr / Scaffold**	**Start (Mb)**	**Gene Annotation**	**$\log_2$FC (Male:Female)**	**$-\log_{10}$ (p-value)**
Family 415					
SapurV1A.4349s0010	4349	0.008	stomatal cytokinesis defective SCD1	-3.6	46.4
SapurV1A.0161s0180	15	16.314	transportin, modifier of snc1 MOS4	-3.4	34.5
SapurV1A.1494s0060	1494	0.062	orcinol O-methyltransferase	-5.4	18.8
SapurV1A.2212s0020	2212	0.02	activating signal cointegrator 1 subunit 3 / RNA helicase DExH14	-4.8	18.5
SapurV1A.0530s0090	15	14.557	wall-associated receptor kinase	-3.4	16.6
SapurV1A.2212s0030	2212	0.023	activating signal cointegrator 1 subunit 3 / RNA helicase DExH14	-5.2	14
SapurV1A.0033s0600	2	11.986	pectinesterase/pectinesterase inhibitor	-5.3	13.5
SapurV1A.1479s0010	1479	0.005	Pre-mRNA-splicing factor ATP-dependent RNA helicase DEAH1	-2.9	13.5
SapurV1A.0582s0010	582	0.028	C3HC zinc finger-like protein NIPA	-2.4	13
SapurV1A.1896s0010	1896	0.001	stomatal cytokinesis defective SCD1	-2.4	13
SapurV1A.0301s0080	15	13.597	extracellular calcium sensing receptor CaS	-1.2	12.1
SapurV1A.2260s0020	2260	0.015	tropinone reductase II	-1.2	11
SapurV1A.0475s0010	15	11.958	poly(A) polymerase 1, nucleotidyltransferase domain	-1.7	8.8
SapurV1A.1332s0070	1332	0.078	PPR containing plant protein	-2.0	8.6
SapurV1A.0665s0030	15	4.5	serine palmitoyltransferase LCB1	-1.6	8.1
SapurV1A.1372s0010	1372	0.025	vacuolar protein sorting-associated protein	-2.3	6.5
SapurV1A.1205s0010	9	3.302	cytochrome P450 family protein CYP81K1	-2.1	6.1
SapurV1A.0665s0090	15	4.385	cyclic nucleotide-gated ion channel protein	-1.4	5.7
SapurV1A.1929s0050	1929	0.021	ISWI chromatin remodeling complex ATPase CHR11	-2.0	5
SapurV1A.2365s0020	2365	0.012	down syndrome critical region protein	-1.7	4.9
SapurV1A.0176s0500	8	6.8	Floral homeotic protein APETALA 1	-1.2	4.9
SapurV1A.0269s0180	10	2.253	galactinol synthase	-1.3	4.8
SapurV1A.0107s0070	3	7.145	importin subunit alpha IMPA-2	-2.3	4.7
SapurV1A.4040s0010	4040	0.005	Di-glucose binding Kinesin motor domain protein	-2.0	4.6
SapurV1A.2266s0020	11	2.678	serine/threonine-kinase	-1.2	4.4
SapurV1A.0685s0220	19	13.327	ankyrin repeat protein	-2.6	4.1
SapurV1A.0542s0040	14	10.924	aminopeptidase	-1.2	4.1
SapurV1A.0122s0480	3	13.047	peroxidase PER7	9.5	14.7
SapurV1A.8306s0010	8306	0.001	polynucleotidyl transferase, ribonuclease H-like	4.5	13.6
SapurV1A.1765s0030	1765	0.028	thylakoid membrane protein	1.7	8.8

Table 6.8
(continued)

***Salix* v1 Gene**	**Chr**	**Start (Mb)**	**Gene Annotation**	**log_2FC**	**$-log_{10}$ (*p*-value)**
SapurV1A.1972s0020	1972	0.021	glutathione S-transferase, tau class GSTU4	4.3	8.6
SapurV1A.0018s0280	16	2.829	SWI/SNF actin-dependent regulator of chromatin A3-like 1	6.0	8.3
SapurV1A.0751s0210	15	3.487	acidic endochitinase	3.1	7.2
SapurV1A.0537s0040	537	0.028	galactoside 2-alpha-L-fucosyltransferase FT1	8.0	6.8
SapurV1A.0189s0060	16	1.217	SWI/SNF actin-dependent regulator of chromatin A3-like 1	5.1	6.4
SapurV1A.0779s0160	17	2.619	peroxidase PER3	2.2	6.3
SapurV1A.0718s0010	15	13.766	WUSCHEL-related homeobox WOX1	1.5	6.1
SapurV1A.0119s0390	3	11.772	plant disease resistance response protein	7.8	5.8
SapurV1A.1656s0010	10	17.545	LRR receptor-like kinase	2.3	5.8
SapurV1A.2759s0010	1	14.783	cytochrome P450 family protein TT7	1.3	5.8
SapurV1A.1762s0030	1762	0.01	AAA-ATPase AATP1	6.3	5.4
SapurV1A.1370s0100	1370	0.069	Myb/SANT-like DNA-binding domain protein	6.7	5.3
SapurV1A.6405s0010	6405	0.001	Cysteine/Histidine-rich C1 domain family protein	2.2	5.1
SapurV1A.1773s0010	16	2.186	patatin phospholipase, sugar-dependent 1	1.6	5.1
SapurV1A.0102s0170	4	10.152	Copia-like polyprotein/retrotransposon, putative	2.5	5.0
SapurV1A.3339s0010	3339	0.011	Naringenin, flavanone 3-hydroxylase F3H	1.4	5.0
SapurV1A.0757s0040	757	0.028	Sedlin, amino-terminal motif protein	6.2	4.9
SapurV1A.0136s0390	3	0.328	lactoylglutathione lyase	2.0	4.9
SapurV1A.0061s0040	10	14.894	plasma membrane intrinsic protein PIP1-1	1.7	4.7
SapurV1A.0186s0040	5	0.823	regulated by cold 2-like protein SRC2	1.6	4.7
SapurV1A.1113s0060	16	10.385	receptor-like kinase	2.5	4.6
SapurV1A.0006s0890	2	2.678	2OG-Fe(II) oxygenase-like oxidoreductase DMR6	1.8	4.5
SapurV1A.0419s0220	6	25.3	sarcosine oxidase	3.5	4.4
SapurV1A.1495s0070	1	15.69	indole-3-acetic acid inducible protein IAA19	1.8	4.4
SapurV1A.0813s0040	813	0.118	tyrosine kinase domain protein	1.6	4.1
SapurV1A.0882s0090	882	0.104	ADP-ribosylation factor(Arf)/Arf-like (Arl) protein	1.6	4.0
Family 423					
SapurV1A.0829s0180	829	0.124	Floral homeotic protein PISTILLATA	-6.6	16.9
SapurV1A.0816s0030	13	4.13	patatin-like phospholipase	-5.1	6.9
SapurV1A.0003s0580	2	14.218	death domain associated protein	-2.1	6.5
SapurV1A.0493s0080	18	0.268	CC-NBS-LRR resistance protein	-1.6	6.5
SapurV1A.0074s0320	15	4.198	zinc finger homeodomain protein 9	-1.3	5.6
SapurV1A.0081s0200	3	16.302	protease inhibitor/seed storage/lipid transfer family protein	-1.3	5.5

Table 6.8
(continued)

***Salix* v1 Gene**	**Chr**	**Start (Mb)**	**Gene Annotation**	**log_2FC**	**$-log_{10}$ (*p*-value)**
SapurV1A.0065s0010	4	17.744	Ras-related protein Rab-2A	-2.7	5.3
SapurV1A.0223s0380	17	16.563	oxidoreductase	-1.5	5.3
SapurV1A.0571s0030	16	21.445	UDP-glycosyltransferase	-1.4	4.8
SapurV1A.0850s0100	850	0.066	IRK-interacting protein, putative	-1.2	4.7
SapurV1A.0797s0010	16	6.034	hypothetical protein	-2.3	4.5
SapurV1A.0056s0020	3	14.077	transmembrane protein	-1.2	4.5
SapurV1A.1017s0040	15	3.317	plastid cell division protein 1	-1.2	4.5
SapurV1A.0007s0430	10	2.889	glyoxal oxidase	-4.4	4.3
SapurV1A.1836s0030	15	7.329	pentatricopeptide repeat-containing protein	2.3	42.1
SapurV1A.1372s0020	1372	0.029	LRR receptor-like serine/threonine-protein kinase	4.7	21.6
SapurV1A.1267s0030	15	10.851	Kinesin-like protein KIN-14K	2.7	17.2
SapurV1A.2619s0010	15	14.394	leucine-rich repeat protein kinase family protein	2.8	12.6
SapurV1A.2205s0010	2205	0.022	LRR receptor-like serine/threonine-protein kinase	4.5	12.2
SapurV1A.1827s0030	15	14.302	receptor-kinase	4.0	10.7
SapurV1A.0719s0090	15	11.616	Kinesin-like protein KIN-14J	2.3	9.3
SapurV1A.0813s0040	813	0.118	tyrosine kinase domain protein	2.3	9.0
SapurV1A.0556s0060	556	0.163	terpene synthase TPS21	2.7	6.7
SapurV1A.1267s0010	15	10.873	Kinesin-like protein KIN-14O	2.3	6.3
SapurV1A.1267s0020	15	10.853	Kinesin-like protein KIN-14J	1.6	5.9
SapurV1A.0480s0200	10	12.534	prefoldin chaperone subunit family protein	1.3	4.9
SapurV1A.0004s1160	2	8.257	THO complex subunit 4B	1.3	4.2
SapurV1A.0200s0470	8	10.02	protein prenylyltransferase SG1	1.6	4.1
SapurV1A.0065s0010	4	17.744	Ras-related protein Rab-2A	-2.7	5.3

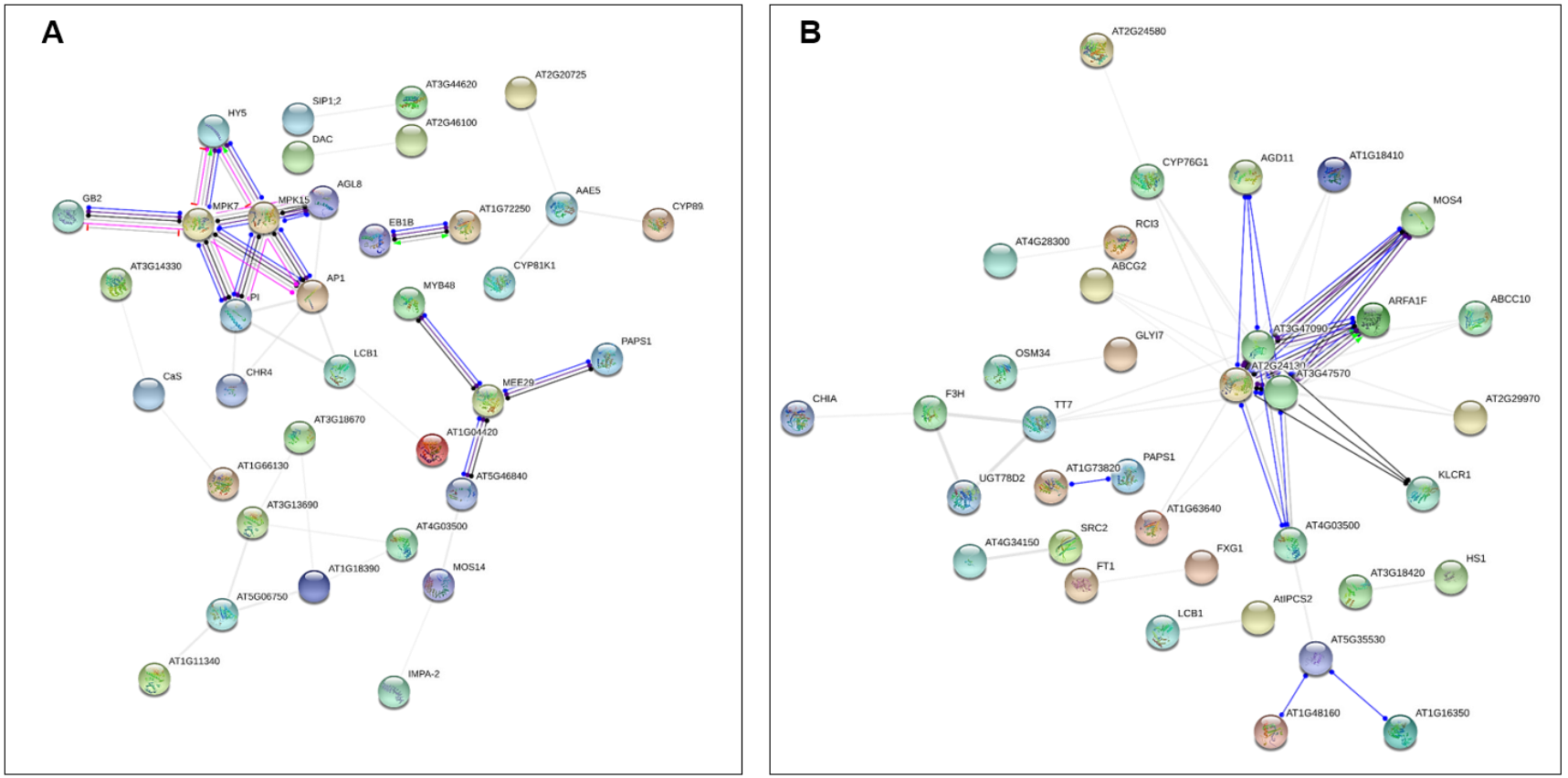

Figure 6.6 STRING network associations of molecular action of **(A)** male-biased and **(B)** female-biased genes (min. conf. = 0.4). Network nodes (circles) represent unique protein-coding genes and edges (lines) represent meaningful protein-protein associations (i.e., jointly contribute to a shared function). Edges are capped by either arrows, perpendicular lines, or dots, and depict positive, negative, and unspecified associations, respectively. Edges are colored according to action types, where black, blue, green, and pink, depict reaction, binding, activation, and posttranslational modification, respectively. Only connected nodes in networks are displayed.

6.4.5 Gene Activation and Silencing

The presence (cpm > 1) or absence (cpm = 0) in the parent and triploid hybrid was compared for each family and Venn diagram analysis was performed. Overall, more genes were silenced than activated in the triploid hybrids, especially for families 415 and 430, which showed nearly five-times the number of genes silenced than activated (Figure 6.7). Family 423 had a greater number of genes activated than the other two triploid families, whereas family 430 had the greatest number of genes silenced. There were no GO-terms enriched for either the activated or silenced gene-sets.

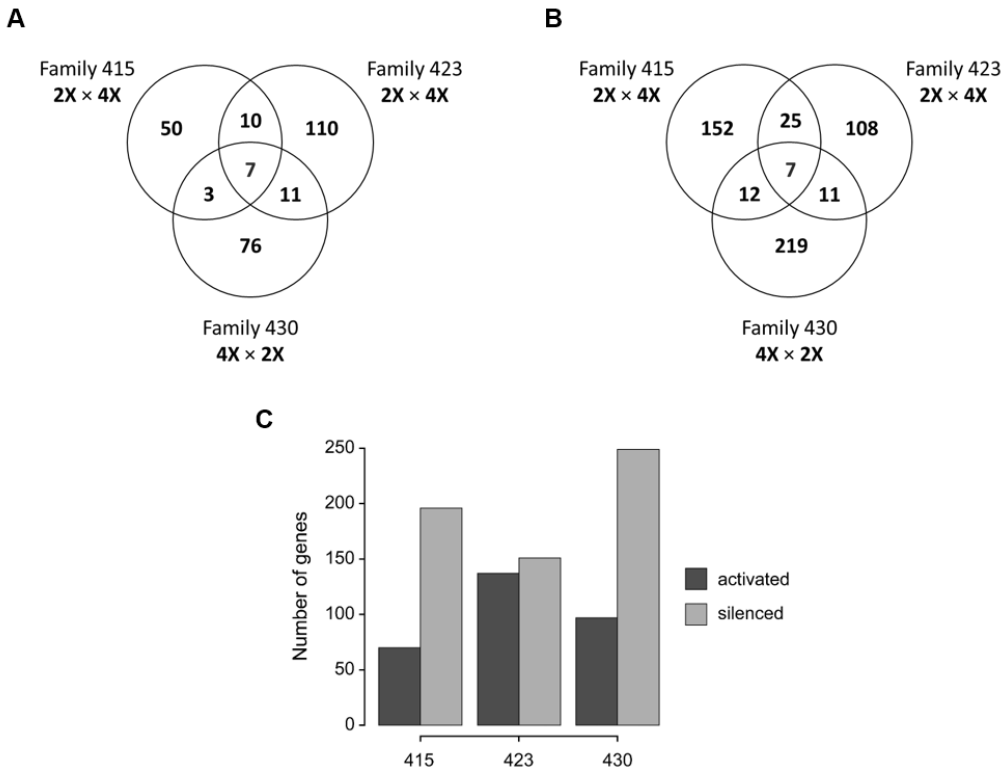

Figure 6.7 Number of genes activated and silenced in triploids and parents. The Venn diagram in **(A)** illustrates genes activated in hybrids and silenced in the parents, whereas the Venn diagram in **(B)** illustrates genes activated in the parents and silenced in the hybrids. The barplot **(C)** summarizes the total number of genes activated (dark grey bars) or silenced (light grey bars) for each family.

6.4.6 Dosage Effects on Gene Expression

To test whether there is a dosage effect on parent alleles in triploid progeny, ASE ratios were compared within and among families. Only extreme deviations from expected dosage ratios ($Pr = 1\times10^{-5}$) were included in the analysis and considered dysregulated. Since it is expected that the triploid hybrid has inherited a single copy of the diploid parent allele and two copies of the tetraploid parent allele, if there was no deviation in expression of the parents alleles in the hybrid, all loci would be represented by a single point at the intersection of expected P_{2X} / P_{4X} difference of logs, $\log_2 (P_{2X} / P_{4X})$ (Figure 6.8).

While the dosage ratios were very similar within in each family, and all genome-wide family means fell within expected ranges, there were significant departures from expected dosage. For dosage ratio outliers, all three triploid families exhibited quite different patterns. Family 415 had high numbers of genes showing up- and down-regulation of the tetraploid parent allele, a majority of which showed *cis* × *trans* regulatory divergence. Family 423 outliers showed up-regulation of the diploid parent allele in the hybrid, as well as high expression levels of both diploid and tetraploid parent alleles in the hybrid. Many outliers with higher diploid and tetraploid ASE in the hybrid were classified ambiguous, and those ratios were not different for the tetraploid parent allele, but a higher diploid ratio had either *trans* or *cis* x *trans* regulatory patterns. All regulatory patterns were represented in family 430, which had the greatest number of unique genes among the families that had dosage ratio outliers from at least one individual. Unlike family 423, there were few genes in family 430 that had genes with greater expression for both diploid and tetraploid parents. In general, common dosage dysregulated genes showed significant enrichment for response to stress, transcription, small molecule activity, and binding activity (Table 6.9).

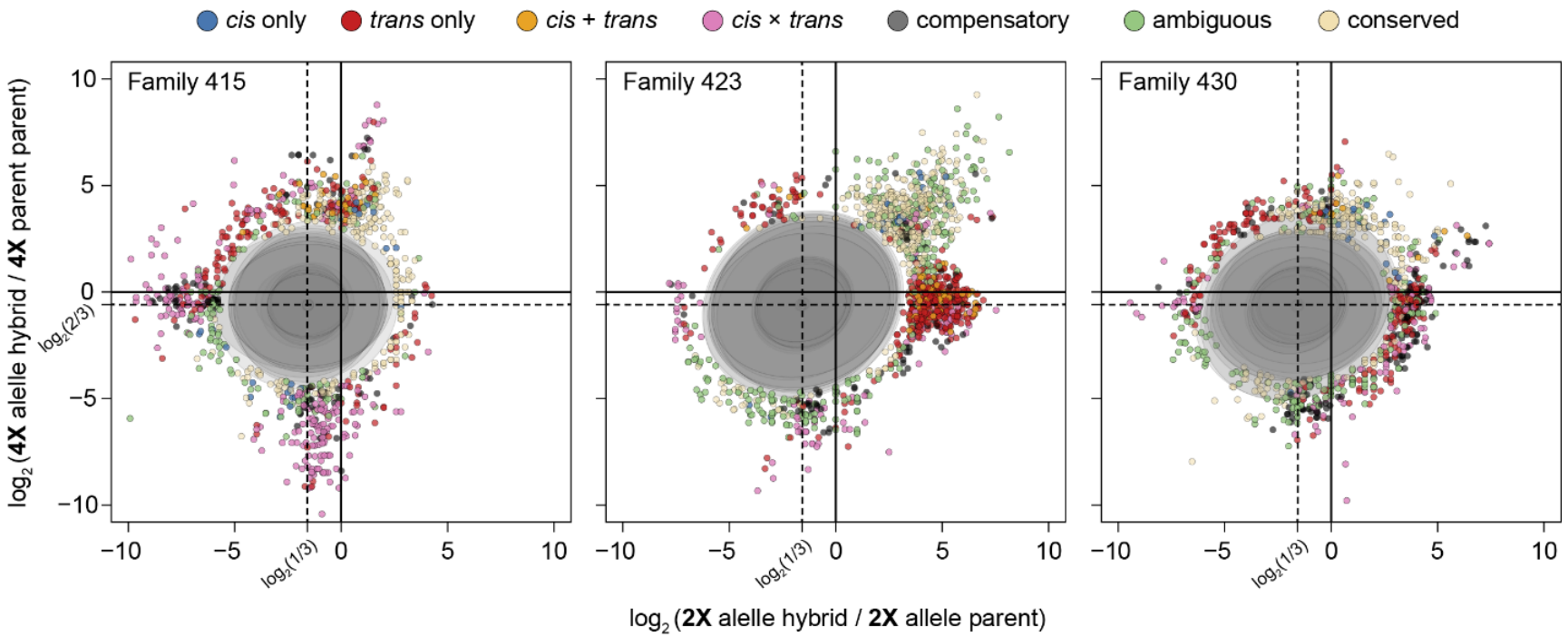

Figure 6.8 Superimposed dosage differential scatterplots of 10 individuals from each of the families 415, 423, and 430 (left to right, respectively). Each point depicts the $\log_2$ ratio of the diploid parent allele in the hybrid and the diploid parent allele in the parent against the $\log_2$ ratio of the tetraploid parent allele in the hybrid and the tetraploid parent allele in the parent. Points are colored according to their regulatory assignment. Ellipses mask a majority of the distribution of $\log_2$ dosage ratios (Pr = 1×10^{-5}), such that points sitting outside ellipses are extreme outliers from expected dosage. Dotted lines at $\log_2(1/3) = -1.585$ and $\log_2(2/3) = -0.585$ represent distribution averages for diploid and tetraploid ratios, and is where the average distribution of dosage ratios are expected to occur.

Table 6.9 Functional enrichments for genes dramatically departed (Pr = 1×10^{-6}) from expected. dosage-based expression.

Pathway ID	Pathway Description	Count	FDR
GO Biological Process			
GO:0009408	response to heat	7	3.5×10^{-05}
GO:0009642	response to light intensity	6	3.5×10^{-05}
GO:0006950	response to stress	18	4.3×10^{-05}
GO:0009266	response to temperature stimulus	9	4.3×10^{-05}
GO:0009628	response to abiotic stimulus	14	4.3×10^{-05}
GO:0050896	response to stimulus	23	5.2×10^{-05}
GO:0010035	response to inorganic substance	10	1.1×10^{-04}
GO:0009644	response to high light intensity	4	1.2×10^{-03}
GO:0042542	response to hydrogen peroxide	4	1.2×10^{-03}
GO:0006979	response to oxidative stress	7	1.4×10^{-03}
GO:0042221	response to chemical	14	6.0×10^{-03}
GO:1901700	response to oxygen-containing compound	10	6.1×10^{-03}
GO:0008152	metabolic process	28	1.6×10^{-02}
GO Molecular Function			
GO:0009055	electron carrier activity	21	5.3×10^{-18}
GO:0031072	heat shock protein binding	13	4.8×10^{-06}
GO Cellular Component			
GO:0005737	cytoplasm	25	3.8×10^{-03}
GO:0005622	intracellular	29	6.2×10^{-03}
GO:0044424	intracellular part	29	6.2×10^{-03}
GO:0005575	cellular component	32	4.2×10^{-02}
GO:0005623	cell	29	4.2×10^{-02}
KEGG Pathways			
K04141	protein processing in endoplasmic reticulum	5	4.6×10^{-05}

6.4.7 Chromosomal Copy Number Variation

The difference in $\log_2 (P_{2X} / P_{4X})$ in the hybrid from respective diploid and tetraploid progenitors can help determine if major departures from expected dosage in the hybrid are a result of copy number variation (CNV) in the tetraploid. For instance, it is expected that triploids inherit one chromosome copy from the diploid parent, and two copies from the tetraploid parent, such that the difference of the hybrid $\log_2 (P_{2X} / P_{4X})$ from the parent is equal to 1. Although these ratios are tetraploid parent informative, aneuploidy in the diploid parent cannot be determined, because at least one diploid parent copy must be present in order to make any inference on chromosomal inheritance patterns in triploid progeny. Further, these ratios are not fully-informative, because any copy number in the tetraploid parent (1/4, 2/4, 3/4, 4/4) can potentially exist in four observable cases in the triploid (4/3, 3/3, 2/3, 1/3). However, the percentage of reads attributable to the diploid parent in the triploid hybrid (i.e., *percent diploid*) can be utilized as a second parameter to rectify overlapping parent-hybrid ratios of different parent and hybrid combinations (Figure 6.1).

While chromosome-wide $\log_2 (P_{2X} / P_{4X})$ expression of the female diploid (*S. purpurea* 94006 and *S. viminalis* 07-MBG-5027) and male tetraploid (*S. miyabeana* 01-200-003) parents showed consistent median values approximately equal to 0, *Salix* chr09 significantly deviated from the expected (Wilcoxon p-value $< 1\times10^{-16}$), with a $\log_2 (P_{2X} / P_{4X})$ of 0.49. This suggests that only three copies of chr09 are present in *S. miyabeana* 01-200-003. This was the case for both families 415 and 423, which were produced from the male tetraploid parent 01-200-003. However, no significant deviations from the expected was observed for the parents of family 430 (*S. miyabeana* 01-200-006 × *S. viminalis* ‘Jorr’).

For family 415, five triploid individuals had a median $\log_2$ (P_{2X} / P_{4X}) parent – hybrid difference of 1.43 and approximately 34% of the reads which could be attributed to the diploid parent for chr09, which is expected, given the male parent was limited to three chr09 copies. The other five individuals had a $\log_2$ (P_{2X} / P_{4X}) difference of 0.45 for chr09, and on average were 50% diploid over all loci for the chromosome. Thus, the latter group in family 415 inherited two of the three tetraploid parent copies of chr09 and the former inherited only one copy. A total of six individuals in family 423 had a $\log_2$ (P_{2X} / P_{4X}) difference of 1.44 and were 33.3% diploid on average for chr09, which is expected if they inherited two copies from the tetraploid, because family 415 and 423 share the same male tetraploid parent. The other four individuals in family 423 had a $\log_2$ (P_{2X} / P_{4X}) difference of 0.47 and were 50% diploid on average, so these individuals only inherited one of the three tetraploid parent copies of chr09.

In addition, it was not uncommon for individuals to possess an additional tetraploid copy of a chromosome and lacking another. For instance, the family 415 individual, 12X-415-031, had a $\log_2$ (P_{2X} / P_{4X}) difference of 1.59 for chr17, but only 25% diploid, which suggests that 12X-415-031 inherited an additional copy of the male tetraploid parent 01-200-003 chr17. Stunningly, the same individual also lacked one copy of the male chr09 ($\log_2$ (P_{2X} / P_{4X}) = 0.45 and 50% diploid) (Figure 6.9A, Figure 6.9B). Another example was for the family 423 individual 12X-423-070 (Figure 6.9C, Figure 6.9D). While 12X-423-070 inherited two copies of chr09 from the tetraploid parent 01-200-003 ($\log_2$ (P_{2X} / P_{4X}) = 1.43 and 33.3% diploid), this individual lacked one copy of the tetraploid parent chr10 ($\log_2$ (P_{2X} / P_{4X}) = 0.0 and 50% diploid), which seems to be spurious, given there was no DNA-Seq or RNA-Seq coverage to indicate that 01-200-003 lacked a copy of chr10.

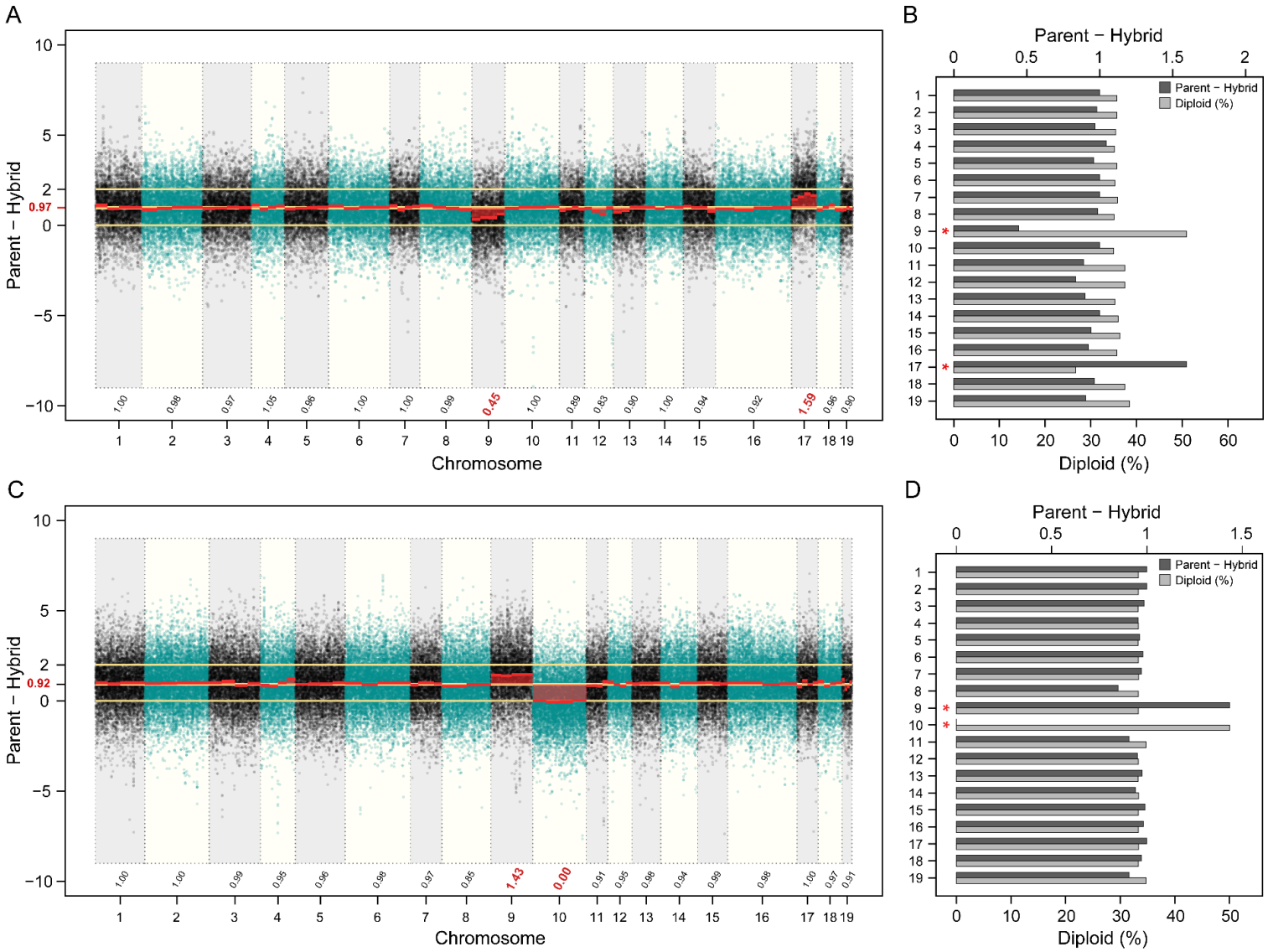

Figure 6.9 Manhattan plot **(A)** chromosome-wide differences of $\log_2$ (P_{2X} / P_{4X}) expression (parent – hybrid) between the family 415 parents (female diploid 94006 and male tetraploid 01-200-003) and the triploid hybrid 12X-415-031. Median parent – hybrid values

Figure 6.9
(continued)

are shown above chromosome identifiers (x-axis). The barplot **(B)** depicts the the median parent–hybrid difference (dark grey bars, scale top x-axis) and the percent expression in the hybrid attributable to the diploid parent allele (light grey bars, scale lower x-axis) by chromosome (y-axis). The Manhattan plot in panel **(C)** and barplot in panel **(D)**, represent the the same analyses, but between the family 423 parents (female diploid 07-MBG-5027 and male tetraploid 01-200-003) and the triploid hybrid 12X-423-070. Red text on x-axes in panels **(A)** and **(D)** correspond to red asterisks on y-axes in panels **(B)** and **(C)**, which denote significant differences (Wilcoxon p-value $< 1\times10^{-16}$).

Unequal inheritance of chr09 in families 415 and 423 was unexpected, yet it permitted a test for genes insensitive to changes in dosage for this chromosome, as well as common genes up- or down-regulated in each group. Three individuals each from families 415 and 423 with a 33% diploid attribution and three each from both families with a 50% diploid attribution for chr09 were compared. Individuals with spurious tetraploid CN (e.g., 12X-415-031 and 12X-423-070) were not included in the analysis. As previously stated, there is a global dosage effect in triploids, irrespective of CN, but dosage sensitive genes – which are most likely to be misexpressed – should show consistent and directional fold-changes. To avoid any buffering effects from the diploid parent (P_{2X}), allele-specific expression of P_{4X} in the parent and hybrid were compared with a binomial exact test in order to reject the null hypothesis that the expression of P_{4X} allele in the triploid hybrid is half (Pr = 0.5) that of P_{4X} allele in the tetraploid parent.

6.4.8 Correlations of Differential Gene Expression with Heterosis for Biomass Traits

Since differences in chromosomal copy numbers in triploids could have drastic phenotypic consequences, Pearson correlations (r^2) were made for genome-wise average diploid (%) and heterosis for important biomass-related growth traits collected in the field and greenhouse. In general, diploid % was positively correlated with heterosis for foliar traits and inversely correlated with heterosis for biomass stem traits (Table 6.10). Diploid % was positively correlated with the field-collected traits FORM, SLA, LFP, LFL, and LFR, and inversely correlated with HT, CDIA, DVOL, DIA, VOL, SA, and LFF. For greenhouse-collected traits, diploid % was positively correlated with SLA only, but inversely correlated with SDW, LDW, RDW, AGB, TOT, HT, MSL, TSL, SA, VOL, and PHE.

Table 6.10 Pearson correlation coefficients (r^2) of heterosis for field- and greenhouse-collected biomass-related traits and the mean percentage of each locus in triploid progeny attributable to the respective diploid parent (i.e., diploid %).

Trait	Time[1]	Trait Class	r^2 - Year 1[2]		r^2 - Year 2	
Field Trial						
HT	1, 2	biomass	−0.82	***	−0.53	**
STNo	1	biomass	−0.41	*	*ns*	
MDIA	1, 2	biomass	−0.47	*	−0.40	*
DIA	1, 2	biomass	−0.54	**	−0.51	**
MSA	1	biomass	−0.41	*	*ns*	
SA	1, 2	biomass	−0.51	**	−0.47	*
VOL	1, 2	biomass	−0.53	**	−0.52	**
DVOL	1, 2	biomass	−0.56	**	−0.58	**
LFL	2	foliar	*ns*		0.45	*
LFP	2	foliar	*ns*		0.50	**
LFR	2	foliar	*ns*		0.42	*
LFF	2	foliar	*ns*		−0.51	**
SLA	1	foliar	0.58	**	*ns*	
CDIA	1, 2	architecture	−0.40	*	−0.65	***
FORM	1, 2	architecture	0.46	*	0.65	***
DEN	2	composition	*ns*		−0.41	*
Greenhouse Trial						
SDW	70	dry wt. biomass	−0.52	**		
LDW	70	dry wt. biomass	−0.57	**		
RDW	70	dry wt. biomass	−0.73	***		
AGB	70	dry wt. biomass	−0.56	**		
TOT	70	dry wt. biomass	−0.65	***		
HT	42, 21-56	biomass	−0.68	***		
MSL	42, 21-56	biomass	−0.60	***		
TSL	42, 21-56	biomass	−0.59	**		
SA	70	biomass	−0.48	**		
VOL	70	biomass	−0.60	***		
SLA	70	foliar	0.31	*		
PHE	11, 13	phenology	−0.52	**		
SPAD	14	physiology	−0.39	*		

[1] Time in years since coppice (field trial) or days after planting (greenhouse trial).

[2] Asterisks ***, **, * denote significant at p-value < 0.001, < 0.01, and < 0.05, respectively.

Field- and greenhouse-collected traits most positively correlated with diploid % were FORM-2 ($r^2 = 0.65$, $p < 0.001$) and SLA ($r^2 = 0.65$, $p < 0.05$), respectively, and most inversely correlated were HT-1 ($r^2 = -0.82$, $p < 0.001$) and RDW ($r^2 = -0.73$, $p < 0.001$), respectively. The only foliar field trait with an inverse association with diploid % was LFF ($r^2 = -0.51$, $p < 0.01$), which is a measure of roundness, or symmetry.

One method of characterizing misexpression is calculating the total deviation in gene expression of genes most commonly expressed across all sample libraries (Kremling et al., 2018). The cumulative expression dysregulation was added to a list of nonadditive and additive inheritance proportions, regulatory divergence proportions, as well as the number of genes expressed, and average normalized expression. The association between these variables were correlated ($p < 0.01$) with heterosis for biomass-collected traits in the field and in the greenhouse (Figure 6.10). Traits included in this analysis are described in Chapter 3 (Table 3.2). Overall, there were stronger associations in the greenhouse trial than the field trial. The total proportion of differentially expressed genes, the cumulative expression dysregulation, and the proportion of differentially expressed genes with additive expression inheritance. Inversely correlated with heterosis for nearly all biomass traits in the greenhouse trial were heterosis for foliar traits was inversely correlated with proportions of additive expression, as well as *cis*- and *cis + trans* divergence. For both the field and greenhouse trials, *trans*-divergence, differential expression, and the proportion of diploid- and tetraploid-parent dominant genes was positively correlated with total volume. Heterosis for hemicellulose content was positively correlated with the proportion of *cis*-divergence and inversely with overdominance proportion. Heterosis for cellulose and lignin were positively and inversely correlated with the proportion of differentially expressed genes with diploid parent dominant expression.

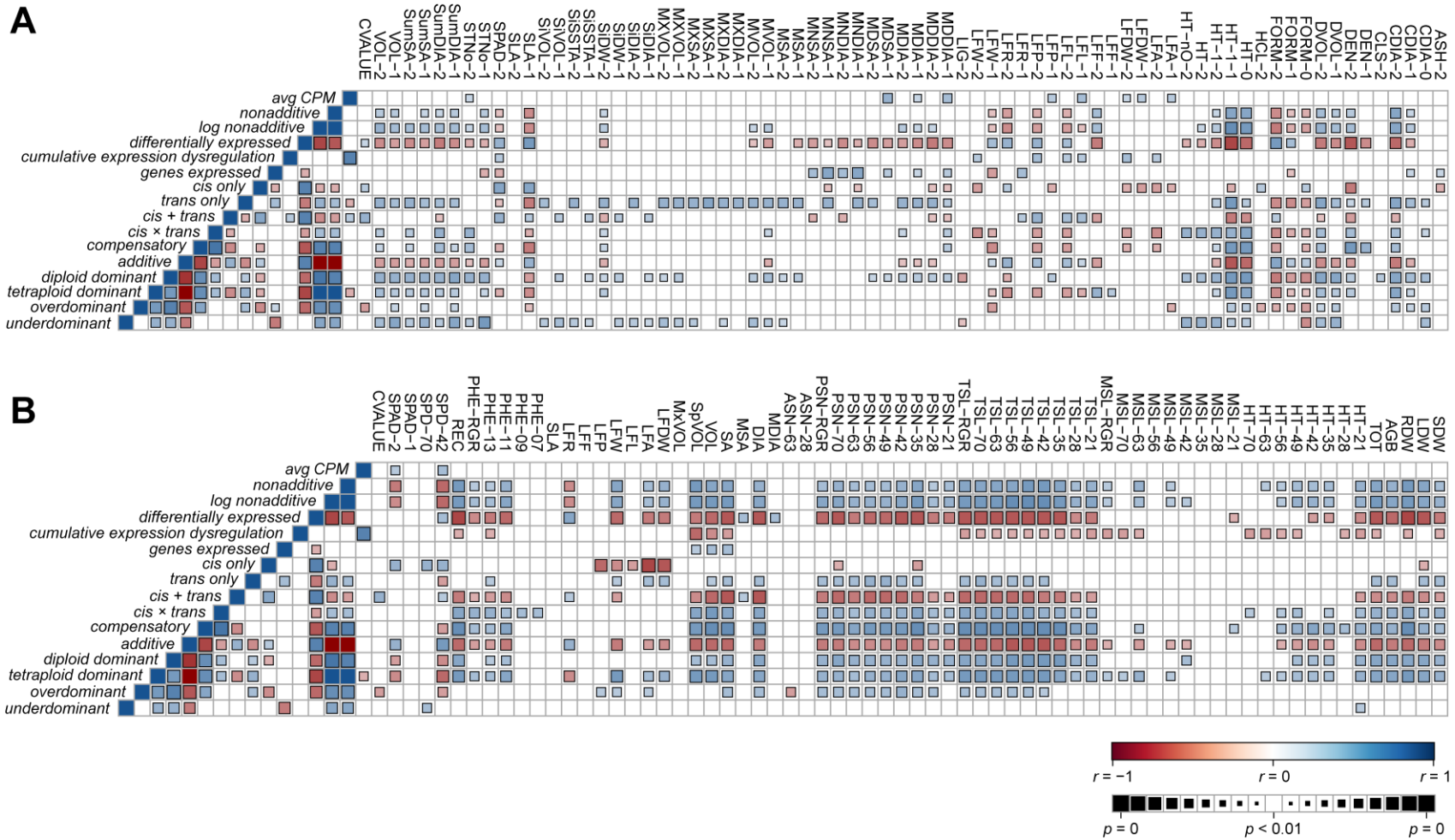

Figure 6.10 Correlation of nonadditive, regulatory divergent, and cumulative expression dysregulation with heterosis for **(A)** field-collected and **(B)** greenhouse-collected traits. Pearson correlation coefficients (*r*), positive correlations are illustrated by filled blue squares and negative correlations by filled-red squares. Non-significant correlations ($p > 0.01$) were left blank. Significance levels (*p*-values) were used to scale the area of each square, such that smaller squares represent correlation coefficients with lower significance and larger squares represent those with greater significance.

Heterosis for SPAD in both trials were inversely correlated with the proportion of nonadditive expression, *trans*- and compensatory regulatory divergence, and the proportion of differentially expressed genes with tetraploid parent dominant inheritance.

The proportion of *cis*-divergence was positively, albeit weakly, correlated with the proportion of differentially expressed genes with additive inheritance, but inversely correlated with the proportion of diploid parent dominant and overdominant expression. The proportion of differentially expressed genes with *trans*-divergent expression was inversely correlated with the proportion of additive inheritance, but positively correlated with dominant and overdominant proportions. Finally, average normalized counts per million (CPM) had the greatest positive association with cumulative dysregulation of gene expression; however, nonadditive inheritance and regulatory divergent class proportions lacked any significant associations with cumulative dysregulation.

6.5 Discussion

6.5.1 Differential Gene Expression is Additive and Nonadditive

Using microarrays of maize, Stupar and Springer (2006) determined that approximately 20% of the genes that were differentially expressed between parents were nonadditively expressed in the hybrid, although very few were above the high parent (overdominant) or below the low parent (underdominant). Swanson and Wagner (2006) found all inheritance categories in the hybrid represented among differentially expressed genes between two inbred parents of maize. In this study, there were very few genes which that differentially expressed between diploid and tetraploid parents which were outside the parental range in the triploid hybrid, especially overdominant genes, which were ~3× less-frequent on average than underdominant genes. In contrast, differentially expressed genes between heterozygous thistle (*C. arvense*)

parents were more frequently overdominant than underdominant in intraspecific hybrids (Bell et al., 2013). Expression-level dominance was most-prominent among both both shoot tip and stem internode tissues of F_1 and F_2 diploid *S. purpurea* families (Carlson et al., 2017, see Chapter 5), and was primarily biased in the direction of the female parent, especially in shoot-tip tissues. Very little additive gene expression was observed in *S. purpurea*, which is a unique result, compared with model crop plants (Guo et al., 2006; Stupar and Springer 2006; Song et al., 2013). Among the triploid *Salix* families investigated in this study, expression-level dominance was prominent, as was established in diploid *S. purpurea*, but the percentage of differential expression attributed to dominance inheritance ranged from 28% to 60%. Reciprocal crosses between *Salix* Sections Vimen and Helix, showed the greatest percentage of dominant expression, which was 50% of those genes expressed differently between diploid and tetraploid parents. Cases of expression-level dominance in polyploid crops have been described in intraspecific thistle (Bell et al., 2013), interspecific coffee (Combes et al., 2015), as well as in allotetraploids of rice (Xu et al., 2014) and Arabidopsis (Shi et al., 2012). Preferential expression is thought to be orchestrated by allelic interactions, which functions to silence one of the parent alleles in a parent-of-origin manner (Chen and Pikaard 1997; Stupar et al., 2007; Donoghue et al., 2014; Baldauf et al., 2016).

Further analysis of allele-specific expression in triploid willow indicates that gene expression variation is associated with both *cis*- and *trans*-regulatory divergence, and that *cis*–*trans* compensatory interactions can account for up to 25% of the variation. Allele-specific expression has been extensively studied in model species, most notably, in interspecific hybrids and allopolyploids of Arabidopsis (Shi et al., 2012) and *Drosophilia* (Landry et al., 2005; Wittkopp et al., 2008a; Wittkopp et al., 2008b; McManus et al., 2010). There is a general trend

that *cis*-regulatory divergence accounts for a greater proportion of expression variation in interspecific hybrids and that *trans*-regulatory divergence is more frequent in intraspecific hybrids (Wittkopp et al., 2004). In hybrids of inbred maize, *cis*-acting variation accounted for most of the divergent expression between parents and was largely attributed to additive expression patterns (Stupar and Springer 2006). Greater sequence divergence was proposed to promote the flexibility of *trans*-factors in their binding to interacting factors and *cis*-elements in *Arabidopsis thaliana* and *A. arenosa* parent alleles (Shi et al., 2012). McManus et al. (2010) hypothesized that greater transgressive inheritance is associated with greater proportions of *cis* × *trans* divergence. In triploid hybrids of *Salix*, greater proportions of overdominant and underdominant expression did coincide with greater proportions of *cis* × *trans* and compensatory regulatory classes. Further, a greater proportion of *cis* + *trans* divergence coincided with a lower proportion of underdominant expression.

6.5.2 Sex-Biased Expression Localizes to the Large, Non-Recombining Sex Determining Region (SDR) on Salix Chromosome 15

In the past decade, there have been numerous reports on sex determination and the evolution of dioecy in plants (Muyle et al., 2017). Sex can have a dramatic influence on gene expression (Coolon et al., 2013; Meilklejohn et al., 2014) and may be one of the most important factors in the evolution of dioecious species. This study provides the first evidence of sexually dimorphic gene expression in triploid willow. From what little is known, it is likely that willows have maintained a female-heterogametic, ZW sex system (Alstrom-Rapaport et al., 1997; Semerikov 2003; Pucholt et al., 2015; Chen et al., 2016; Pucholt et al., 2017; Zhou et al., 2018), in contrast to the XY system of poplars (Yin et al., 2008; Geraldes et al., 2015). Although the genomes of *S. purpurea* and closely-related *P. trichocarpa* are rather colinear (Rodgers-Melnick

et al., 2012), the sex determining regions (SDR) are located on chr19 and chr15, respectively. Based on the recent characterization of the large non-recombining region on *S. purpurea* chr15 (Zhou et al., 2018), it is likely that chr19 is the ancestral sex determining chromosome in the Salicaceae.

A majority of the sex determination research in the Salicaceae has focused on identifying QTL linked to sex in experimental mapping populations (Carlson et al., 2018; see Chapter 4). Yet, only a handful of studies in willow have used RNA-Seq for inference - of which little is still known about the underlying regulators in sex determination. For instance, RNA-sequencing of catkin tissue in *S. suchowensis* (Liu et al., 2013) and *S. viminalis* (Pucholt et al., 2017) identified numerous genes differentially expressed by sex, but failed to establish any biological links among those with significant associations. However, using shoot tip tissues intraspecific F_1 *S. purpurea*, which likely contained differentiating floral buds, Carlson et al. (2017) found that the most significant differentially expressed genes were positioned within the non-recombining, SDR on *Salix* chr15. Significantly, many genes differentially expressed in intraspecific F_1 *S. purpurea* were also identified as differentially expressed between female and male triploid individuals in families 415 and 423. There were more differentially expressed genes in common with intraspecific *S. purpurea* within family 415, which is not surprising, because they share the common female diploid parent 94006, and it is likely that some of the top female-biased genes represent W-specific segments of chr15.

The most significant ($\log_2$ (male/female) = −3.6, $-\log_{10}$ (*p*-value) = 46.4) was for the *Salix* gene model, SapurV1A.4349s0010, which is annotated as a stomatal cytokinesis-defective 1 (SCD1) protein. The gene is located on a *S. purpurea* v1 scaffold and represents an unassembled W haplotype of chr15. The function of SCD1 is not well-known, but is implicated

in cytokinesis and polarized cell expansion in epidermal cells as well as flower morphogenesis (Falbel et al., 2003; Korasick et al., 2010; Mayers et al., 2017). However, not all genes differentially expressed between sexes were located on chr15. Two encode for the floral homeotic proteins, *PISTILLATA* (PI) and *APATELLA 1* (AP1). PI was most significantly associated with sex in family 423. Both genes are known to physically interact, as PI can form a heterodimer with either AP1 or AP3, functioning principally to regulate floral organ identity in differentiating primordial tissues (Bartlett et al., 2016). These genes are likely major players in the determination of sex in *Salix*, but differential expression does not necessarily imply causation. It is possible that their expression, as well as other autosomal genes, are modulated by a sex determining factor within the non-recombining SDR on chr15. Both genes are subject to silencing by miRNA (Zhao et al., 2009), which are known to be quite dosage-sensitive, so it is plausible that dosage of the W haplotype could influence global miRNA accumulation in females.

6.5.3 Global Dosage Balance with Local Sensitivities

Dosage in all three triploid families appears to behave in an extraordinarily additive manner, irrespective the number of parent copies inherited. However, there were a handful of genes that did depart from expected dosage in triploids, most notably, those coding for heat shock proteins. In this study, genes encoding for heat shock proteins displayed greater expression in individuals with normal chr09 copies, whereas those null a tetraploid parent copy had greater expression of stress- or senescence-associated genes. Overall, there were greater proportions of loci showing *cis* × *trans* and compensatory regulatory patterns in family 415 and 423 individuals null one tetraploid parent copy (e.g., chr02, chr09, chr10, and chr17), or a greater average diploid %. While the quantity of a translation product (protein subunit) may impact the assembly of a

particular complex, the mere involvement in a complex can also impact protein stability (Veitia et al., 2008). It may be that null mutations in metabolic functions are tolerated in a heterozygous state, but only weak, loss-of-function, dosage-sensitive genes can survive negative selection as heterozygotes (Birchler and Veitia, 2010). The balance of regulatory hierarchies, i.e., dosage balance (Birchler et al., 2005), are sensitive to gene dosage and changes in individual components can have an effect on the phenotype. In macromolecular complexes, dosage balance is essential, because partial aneuploidy of a dosage-sensitive gene can change the stoichiometry of the complexes and lead to fitness defects (Veitia et al., 2008). In maize, greater proportions of nonadditive expression was observed in triploid and tetraploid hybrids with genome dosage effects (Guo et al., 1996; Auger et al., 2005; Birchler et al., 2005; Riddle et al., 2010).

Previous gene expression studies in inbred and outcrossing species have regularly pooled F_1 progeny libraries prior to sequencing. While this is likely not an issue for inbred crops, these results show that pooled RNA-Seq can underestimate factors contributing to the inheritance of gene expression in highly heterozygous species, especially for families derived from natural polyploids. For instance, without sequencing individual libraries, it would be impossible to detect null chromosomes of polyploid progenitors in the F_1 based on pooled RNA-Seq data alone, which could distort assumptions about the evolution of gene expression inferred from inheritance and regulatory assignments. Even if the expected ploidy in the hybrid is based on chromosome counts or DNA-Seq of the parents, there may not be equal inheritance, and binomial tests for ASE between the parents and the hybrid would be incorrect if based on a fixed probability estimate. Thus, prior to tests for ASE, a simple adjustment could be made, which would first require that each chromosome (or scaffold) be tested independently. Utilizing median

fold-changes in the parents and the percentage of reads in the hybrid attributable to the diploid parent, a probability of success under the null could be properly assigned.

Beyond the fact that the parents in this study were unimproved and highly-heterozygous, it is possible that CNV or aneuploidy can help explain some of the variation in heterosis observed within and among triploid families. In aneuploid studies, changing numerous chromosome segments can alter quantitative characters (Guo and Birchler, 1994). Here, genome-wide averages of ASE attributable to the diploid parent, i.e., diploid %, in triploids was inversely correlated with heterosis for important stem growth traits (e.g., total harvestable biomass), but positively correlated with heterosis for foliar traits. The dosage balance hypothesis, outlined by Birchler (2005), may very well apply to slight deviations in the global inheritance of parent ASE or major differences in chromosomal copy number (i.e., aneuploidy), as was observed for *Salix* chr09 aberrations in families 415 and 423. Genetic mapping of complex biomass-related traits in F_2 *S. purpurea* (Carlson et al., 2018; Chapter 4) identified QTL on chr09 for leaf length (LFL), leaf perimeter (LFP), and specific leaf area (SLA), so positive correlations between diploid % and foliar traits could indicate a dosage sensitivity of genes controlling the variation for these traits.

6.5.4 Nonadditive Gene Expression Correlates with Nonadditive Phenotypic Expression

One of the major challenges in molecular genetics is disentangling the relationship of transcriptome-wide expression patterns to phenotypic effects (Birchler 2007). Rather than concentrating on the terminologies of heterosis models, e.g., dominance, overdominance or pseudo-overdominance, Birchler et al. (2010) described a progression to a more quantitative and interactive or network-oriented framework for dissecting the phenomenon of heterosis. Utilizing DNA-Seq and RNA-Seq to unravel the underlying regulatory architecture of differential

expression, the current understanding of heterosis in high-yielding triploid hybrids of willow has improved. This work shows the proportion of genes differentially expressed between diploid and tetraploid parents attributable to nonadditive gene expression in the triploid hybrid (namely expression-level dominance) was positively correlated with heterosis for biomass yield as well as biomass-related growth traits collected in the greenhouse and in the field. In addition, this study corroborates some of the key findings reported in Kremling et al. (2018), such that cumulative expression dysregulation is inversely correlated with heterosis for biomass and that individuals with greater absolute expression tended to display greater levels of dysregulation.

Importantly, tetraploid parent dominant genes among triploid hybrids were enriched for the following pathways: phenylpropanoid biosynthesis, cyanoamino acid metabolism, biosynthesis of secondary metabolites, and starch and sucrose metabolism. Some of the most interesting tetraploid parent dominant genes identified in this study were those annotated as uridine diphosphate (UDP) glycosyltransferases (UGTs). UGTs catalyze the transfer of sugars to a wide range of acceptor molecules, including plant hormones, and all classes of plant secondary metabolites (Ross et al., 2001). Recent linkage mapping in a full-sib F_2 *S. purpurea* family identified QTL associated with many important traits for biomass production (Carlson et al., 2018; see Chapter 4), including biomass growth, wood chemical composition, as well as for foliar, physiological, and pathology-related traits. Further analysis of this candidate gene set, with regards to their significance in overlapping support intervals from mapping experiments or regulatory patterns in other high-yielding triploid hybrid individuals will prove useful in the genetic improvement of shrub willow as a bioenergy crop.

6.6 Conclusion

This study provides evidence that nonadditive gene expression correlates with nonadditive phenotypic expression in triploid hybrids of willow. Expression-level dominance was most correlated with heterosis for biomass yield and stem volume and were highly enriched for important processes involved in starch and sucrose metabolism. There was a global dosage effect of parent alleles in triploid hybrids, with few that were insensitive to copy number variation. Finally, this study reports the first evidence of sexually dimorphic gene expression in triploid willow. It is vital that we apply our ever-improving understanding of heterosis from studies of well-characterized diploid crop species, such as maize, tomato, and rice to the improvement of yield and biomass quality of undomesticated crops, including willow and poplar, which are being developed to provide sustainable sources of lignocellulosic biomass for bioenergy, biofuels, and bioproducts. Additional characterization of the genomic basis of heterosis in closely related genera or more diverse crosses could be useful in determining the evolutionary benefits of wide hybridization and polyploidy in *Salix*.

6.7 REFERENCES

Alström-Rapaport C, Lascoux M, Gullberg U. 1997. Sex determination and sex ratio in the dioecious shrub *Salix viminalis* L. *Theoretical and Applied Genetics*, 94: 493–497.

Auger DL, Gray AD, Ream TS, Kato A, Coe EH, Birchler JA. 2005. Nonadditive gene expression in diploid and triploid hybrids of maize. *Genetics*, 169: 389–397.

Bartlett M, Thompson B, Brabazon H, Del Gizzi R, Zhang T, Whipple C. Evolutionary dynamics of floral homeotic transcription factor protein-preotein interactions. *Molecular Biology and Evolution*, 33: 1486–1501.

Baldauf JA, Marcon C, Paschold A, Hochholdinger F. 2016. Nonsyntenic genes drive tissue-specific dynamics of differential, nonadditive, and allelic expression patterns in maize hybrids. *Plant Physiology*, 171: 1144–1155.

Bell GD, Kane NC, Rieseberg LH, Adams KL. 2013. RNA-seq analysis of allele-specific expression, hybrid effects, and regulatory divergence in hybrids compared with their parents from natural populations. *Genome Biology and Evolution*, 5: 1309–1323.

Berlin S, Lagercrantz U, von Arnold S, Ost T, Ronnberg-Wastljung A. 2010. High-density linkage mapping and evolution of paralogs and orthologs in *Salix* and *Populus*. *BMC Genomics*, 11(1): 129.

Birchler JA, Auger DL, Riddle NC. 2003. In search of the molecular basis of heterosis. *The Plant Cell*, 15: 2236–2239.

Birchler JA, Riddle NC, Auger DL, Veitia RA. 2005. Dosage balance in gene regulation: biological implications. *Trends in Genetics*, 21: 219–226.

Birchler JA, Veitia RA. 2007. The gene balance hypothesis: From classical genetics to modern genomics. *The Plant Cell*, 19: 395–402.

Birchler, JA. 2010. Heterosis. *The Plant Cell*, 22: 2105–2112.

Birchler JA, Veitia RNA. 2010. The gene balance hypothesis: Implications for gene regulation, quantitative traits and evolution. *New Phytologist*, 186: 54–62.

Bonosi L, Ghelardini L, Weih M. 2013. Towards making willows potential bio-resources in the South: Northern *Salix* hybrids can cope with warm and dry climate when irrigated. *Biomass and Bioenergy*, 51: 136–144.

Carlson CH, Smart LB. 2016. Electrical capacitance as a predictor of root dry weight in shrub willow (*Salix*; Salicaceae) parents and progeny. *Applications in Plant Science*, 4: e1600031.

Carlson CH, Choi Y, Chan AP, Serapiglia MJ, Town CD, Smart LB. 2017. Dominance and sexual dimorphism pervade the *Salix purpurea* L. transcriptome. *Genome Biology and Evolution*, 9: 2377–2394.

Carlson CH, Gouker FE, Evans LM, DiFazio SP, Crowell CR, Smart DC, Smart LB. 2018. Joint linkage and association mapping of complex traits in shrub willow (*Salix purpurea* L.). *Annals of Botany*, submitted.

Chen CZ, Pikaard CS. 1997. Transcriptional analysis of nucleolar dominance in polyploid plants: Biased expression/silencing of progenitor rRNA genes is developmentally regulated in *Brassica*. *Proceedings of the National Acadamy of Sciences of the United States of America*, 94: 3442–3447.

Chen Y, Wang T, Fang L, Li X, Yin T. 2016. Confirmation of single-locus sex determination and female heterogamety in willow based on linkage analysis. *PLoS ONE*, 11: e0147671.

Combes MC, Hueber Y, Dereeper A, Rialle S, Herrera JC, Lashermes P. 2015. Regulatory divergence between parental alleles determines gene expression patterns in hybrids. *Molecular Biology and Evolution,* 7: 1110–1121.

Coolon JD, Webb W, Wittkopp PJ. 2013. Sex-specific effects of *cis*-regulatory variants in *Drosophila melanogaster*. *Genetics,* 195: 1419–1422.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43: 491–498.

Donoghue MT, Fort A, Clifton R, Zhang X, McKeown PC, Voight-Zielinski ML, Borevitz JO, Spillane C. 2014. C(m)CGG methylation-independent parent-of-origin effects on genome-wide transcript levels in isogenic reciprocal F1 triploid plants. *DNA Research,* 21: 141–151.

Duvick DN. 1999. Heterosis: feeding people and protecting natural resources. In: J. G. Coors and S. Pandey (Eds.) The Genetics and Exploitation of Heterosis in Crops. Madison, WI, American Society of Agronomy; Crop Science Society of America; Soil Science Society of America, pp. 19–29.

East EM. 1936. Heterosis. *Genetics*, 21: 375–397.

Fabio ES, Kemanian AR, Montes F, Miller RO, Smart LB. 2017. A mixed model approach for evaluating yield improvements in interspecific hybrids of shrub willow, a dedicated bioenergy crop. *Industrial Crops and Production,* 96: 57–70.

Falbel TG, Koch LM, Nadeau JA, Segui-Simarro JM, Sack FD, Bednarek SY. 2003. SCD1 is required for cell cytokinesis and polarized cell expansion in *Arabidopsis thaliana*. *Development*, 130: 4011–4024.

Geraldes A, Hefer CA, Capron A, Kolosova N, Martinez-Nuñez F, Soolanayakanahally RY, Stanton B, Guy RD, Mansfield SD, Douglas CJ, Cronk QC. 2015. Recent Y chromosome divergence despite ancient origin of dioecy in poplars (*Populus*). *Molecular Ecology*, 24: 3243–3256.

Goff SA. 2011. A unifying theory for general multigenic heterosis: energy efficiency, protein metabolism, and implications for molecular breeding. *New Phytologist*, 189: 923–37.

Gunter LE, Roberts GT, Lee K, Larimer FW, Tuskan GA. 2003. The development of two flanking SCAR markers linked to a sex determination locus in *Salix viminalis* L. *Journal of Heredity*, 94: 185–189.

Guo M, Birchler JA. 1994. *Trans*-acting dosage effects on the expression of model gene systems in maize aneuploids. *Science*, 266: 1999-2002.

Guo M, Davis D, Birchler JA. 1996. Dosage effects on gene expression in a maize ploidy series. *Genetics*, 142: 1349–1355.

Guo M, Rupe MA, Zinselmeier C, Habben J, Bowen BA, Smith OS. 2004. Allelic variation of gene expression in maize hybrids. *The Plant Cell*, 16: 1707–1716.

Guo M, Rupe MA, Yang X, Crasta O, Zinselmeier C, Smith OS, Bowen B. 2006. Genome-wide transcript analysis of maize hybrids: allelic additive gene expression and yield heterosis. *Theoretical and Applied Genetics*, 113: 831–845.

Hallingback HR, Fogelqvist J, Powers SJ, Turrion-Gomez J, Rossiter R, Amey J, Martin T, Weih M, Gyllenstrand N, Karp A, Lagercrantz U, Hanley SJ, Berlin S, Ronnberg-Wastljung A-C. 2016. Association mapping in *Salix viminalis* L. (Salicaceae) - identification of candidate genes associated with growth and phenology. *GCB Bioenergy*, 8: 670–685.

Hanley SJ, Karp A. 2013. Genetic strategies for dissecting complex traits in biomass willows (*Salix* spp.). *Tree Physiology,* 34: 1167–1180.

He GM, Zhu XP, Elling AA, Chen LB, Wang XF, Guo L, Liang MZ, He H, Zhang HY, Chen FF, Qi YJ, Chen RS, Deng, XW. 2010. Global epigenetic and transcriptional trends among two rice subspecies and their reciprocal hybrids. *The Plant Cell,* 22: 17–33.

Zhuang Y, Adams KL. 2007. Extensive allelic variation in gene expression in *Populus* F_1 hybrids. *Genetics*, 177: 1987–1996.

Korasick DA, McMichael C, Walker KA, Anderson JC, Bednarek SY, Hesse A. Novel functions of Stomatal Cytokinesis-Defective 1 (SCD1) in innate immune responses against bacteria. *The Journal of Biological Chemistry*, 285: 23342–23350.

Kremling KAG, Chen S-Y, Su M-H, Lepak NK, Romay C, Swarts KL, Lu F, Lorant A, Bradbury PJ, Buckler ES. Dysregulation of expression correlates with rare allele burden and fitness loss in maize. *Nature*, 555: 520–523.

Landry CR, Wittkopp PJ, Taubes CH, Ranz JM, Clark AG, Hartl DL. 2005. Compensatory *cis-trans* evolution and the dysregulation of gene expression in interspecific hybrids of *Drosophilia*. *Genetics*, 171: 1813–1822.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics,* 25: 1754–1760.

Liu J, Yin T, Ye N, Chen Y, Yin T, Liu M, Hassani D. 2013. Transcriptome analysis of the differentially expressed genes in the male and female shrub willows (*Salix suchowensis*). *PLoS ONE,* 8: e60181.

Mayers JR, Hu T, Wang C, Cardenas JJ, Ran Y, Pan J, Bednarek SY. 2017. SCD1 and SCD2 form a complex with exocyst and RabE1 in exocytosis and cytokinesis. *The Plant Cell*, 29: 2610–2625.

McManus CJ, Coolon JD, Duff MO, Eipper-Mains J, Graveley BR, Wittkopp PJ. 2010. Regulatory divergence in *Drosophila* revealed by mRNA-seq. *Genome Research,* 20: 816–825.

Meiklejohn CD, Coolon JD, Hartl DL, Wittkopp PJ. 2014. The roles of *cis*- and *trans*-regulation in the evolution of regulatory incompatibilities and sexually dimorphic gene expression. *Genome Research,* 24: 84–95.

Muyle A, Shearn R, Marais GAB. 2017. The evolution of sex chromosomes and dosage compensation in plants. *Genome Biology and Evolution*, 9: 627–645.

Ni ZF, Kim ED, Ha MS, Lackey E, Liu JX, Zhang YR, Sun QX, Chen ZJ. 2009. Altered circadian rhythms regulate growth vigour in hybrids and allopolyploids. *Nature*, 457: 327.

Pucholt P, Rönnberg-Wästljung A-C, Berlin S. 2015. Single locus sex determination and female heterogamety in the basket willow (*Salix viminalis* L.). *Heredity*, 114: 575–583.

Pucholt P, Wright AE, Conze LL, Mank JE, Berlin S. 2017. Recent sex chromosome divergence despite ancient dioecy in the willow *Salix viminalis*. *Molecular Biology and Evolution,* 34: 1991–2001.

R Core Team. 2015. *R: A language and environment for statistical computing.* Vienna, Australia: R Foundation for Statistical Computing.

Riddle NC, Jiang HM, An LL, Doerge RW, Birchler JA. 2010. Gene expression analysis at the intersection of ploidy and hybridity in maize. *Theoretical and Applied Genetics,* 120: 341–353.

Ross J, Li Y, Lim E-K, Bowles DJ. 2001. Higher plant glycosyltransferases. *Genome Biology*, 2: reviews3004.1–3004.6.

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics,* 26: 139–140.

Rodgers-Melnick E, Mane SP, Dharmawardhana P, Slavov GT, Crasta OR, Strauss SH, Brunner AM, DiFazio SP. 2012. Contrasting patterns of evolution following whole genome versus tandem duplication events in *Populus*. *Genome Research*, 22: 95–105.

Semerikov V, Lagercrantz U, Tsarouhas V, Rönnberg-Wästljung Alström-Rapaport Lascoux M. 2003. Genetic mapping of sex-linked markers in *Salix viminalis* L. *Heredity*, 91: 293–299.

Serapiglia MJ, Cameron KD, Stipanovic AJ, Smart LB. 2012. Correlations of expression of cell wall biosynthesis genes with variation in biomass composition in shrub willow (*Salix* spp.) biomass crops. *Tree Genetics and Genomes,* 8: 775–788.

Shi X, Ng DWK, Zhang C, Comai L, Ye W, Chen J. 2012. *Cis*- and *trans*-regulatory divergence between progenitor species determines gene-expression novelty in *Arabidopsis* allopolyploids. *Nature Communications,* 3: 950.

Smart LB, Cameron KD. 2008. Genetic improvement of willow (*Salix* spp.) as a dedicated bioenergy crop. In: Vermerris W, editor. Genetic Improvement of Bioenergy Crops: Springer New York. pp. 377–396.

Song G, Guo Z, Liu Z, Cheng Q, Qu X, Chen R, Jiang D, Liu C, Wang W, Sun Y, Zhang L, Zhu Y, Yang D. 2013. Global RNA sequencing reveals that genotype-dependent allele-specific expression contributes to differential expression in rice F_1 hybrids. *BMC Plant Biology*, 13: 221.

Stupar RM, Springer NM. 2006. *Cis*-transcriptional variation in maize inbred lines B73 and Mo17 leads to additive expression patterns in the F1 hybrid. *Genetics*, 173: 2199–2210.

Springer NM, Stupar RM. 2007a. Allele-specific expression patterns reveal biases and embryo-specific parent-of-origin effects in hybrid maize. *Plant Cell*, 19: 2391–2402.

Stupar RM, Hermanson PJ, Springer NM. 2007b. Nonadditive expression and parent-of-origin effects identified by microarray and allele-specific expression profiling of maize endosperm. *Plant Physiology*, 145: 411–425.

Stupar RM, Gardiner JM, Oldre AG, Haun WJ, Chandler VL, Springer NM. 2008. Gene expression analyses in maize inbreds and hybrids with varying levels of heterosis. *BMC Plant Biology*, 8: 33.

Swanson-Wagner RA, Jia Y, DeCook R, Borsuk LA, Nettleton D, Schnable PS. 2006. All possible modes of gene action are observed in a global comparison of gene expression in a maize F_1 hybrid and its inbred parents. *Proceedings of the National Academy of Sciences U. S. A.*, 103: 6805–6510.

Veitia RA, Bottani S, Birchler JA. 2008. Cellular reactions to gene dosage balance: Genomic, transcriptomic and proteomic effects. *Trends in Genetics*, 24: 390–397.

Wei G, Tao Y, Liu GZ, Chen C, Luo RY, Xia HA, Gan Q, Zeng HP, Lu ZK, Han YN, Li XB, Song GS, Zhai HL, Peng YG, Li DY, Xu HL, Wei XL, Cao ML, Deng HF, Xin YY, Fu XQ, Yuan LP, Yu J, Zhu Z, Zhu LH. 2009. A transcriptomic analysis of superhybrid rice LYP9 and its parents. *Proceedings of the National Academy of Sciences U. S. A.*, 106: 7695–7701.

Wittkopp PJ, Haerum BK, Clark AG. 2004. Evolutionary changes in *cis* and *trans* gene regulation. *Nature*, 430: 85–88.

Wittkopp PJ, Haerum BK, Clark AG. 2008a. Regulatory changes underlying expression differences within and between *Drosophila* species. *Nature Genetics*, 40: 346–350.

Wittkopp PJ, Haerum BK, Clark AG. 2008b. Independent effects of *cis*- and *trans*-regulatory variation on gene expression in *Drosophila melanogaster*. *Genetics*, 178: 1831–1835.

Wittkopp PJ, Kalay G. 2011. *Cis*-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nature Review Genetics*, 13: 59–69.

Wray G. 2007. Evolutionary significance of *cis*-regulatory mutations. *Nature Review Genetics*, 8: 206–216.

Xu C, Bai Y, Lin Y, Zhao N, Hu L, Gong Z, Wendel JF, Liu B. 2014. Genome-wide disruption of gene expression in allopolyploids but not hybrids of rice subspecies. *Molecular Biology and Evolution*, 31: 1066–1076.

Yin T, DiFazio SP, Gunter LE, Zhang X, Sewell MM, Woolbright SA, Allan GJ, Kelleher CT, Douglas CJ, Wang M, Tuskan GA. 2008. Genome structure and emerging evidence of an incipient sex chromosome in *Populus*. *Genome Research*, 18: 422–430.

Zhao L, Kim Y, Dinh TT, Chen X. 2007. miR172 regulates stem cell fate and defines the inner boundary of *APETALA3* and *PISTILLATA* expression domain in Arabidopsis floral meristems. *The Plant Journal*, 51: 840–849.

Zhou R, Macaya-Sanz D, Rodgers-Melnick E, Carlson CH, Gouker FE, Evans LM, Schmutz J, Jenkins JW, Yan J, Tuskan GA, Smart LB, DiFazio SP. 2018. Characterization of a large sexually dimorphic genome interval in *Salix purpurea* L. (Salicaceae). *Molecular Genetics and Genomics,* DOI: 10.1007/s00438-018-1473-y.

CHAPTER 7

FUTURE PERSPECTIVES

7.1 Future Perspectives

Shrub willow (*Salix* spp.) in the subgenus Vetrix are an extremely diverse group of fast-growing woody perennial shrubs. All willows are dioecious and primarily pollinated by insects (entomophilous) in the early spring. The inflorescence of willows can be quite dramatic, especially for males, which produce a prolific amount of pollen, thus, females may be wind-pollinated (anemophilous) on dry, windy days as well. Having two mechanisms for pollen dispersal across the genus *Salix* have likely contributed to the tremendous distribution of the genus around the world, which touts a range from the arctic to the subtropics (Kuzovkina et al., 2008). Commonly found along streambanks and riverbeds, willows flourish in riparian and marginal habitats, as seeds that have thin or no seed coats, and germinate shortly after dispersal. In addition to being an outcrosser, shrub willows clonally reproduce via vegetative propagation. In a natural setting, stem segments buried in topsoil initiate adventitious roots from axial buds along sylleptic branches. Clonal reproduction can limit genetic diversity in natural stands, especially for isolated and naturalized populations (Lin et al., 2009; Lauron-Moreau et al., 2015). Nevertheless, hybridization is a feature of willows and have been shown to hybridize between distinct populations and species in a natural setting (Hardig et al., 2000). The number of species is estimated to be approximately 350 (Argus 1997). No other genus in the Salicaceae comprise such a species diversity as that of the genus *Salix*. For instance, closely-related poplar (*Populus* spp.) shares a common whole-genome duplication event with willow, yet only 23 species have

been characterized, and while the base ploidy of all poplars is diploid (n = 19), the ploidy levels of willow ranges from diploid to dodecaploid, the latter most prevalent in dwarf shrub species.

It is this diversity and adaptive plasticity that led to the use of shrub willow as a bioenergy crop. Willows are bred to be fast growing, genetically diverse, disease and pest-resistant, and able to grow on marginal lands without competing with food crops (Bonosi et al., 2013). European shrub willow breeding programs have historically focused on *S. viminalis* (common osier willow, Sect. Vimen), whereas North American breeding programs focused on *S. purpurea* (purple osier willow, Sect. Helix), which was naturalized by early European settlers. Traditional breeding methods have been used in shrub willow improvement, and involve the development of families from superior parents, phenotyping and preliminary selection in the field, then final selection based on large-scale multi-environment yield trial data (Smart and Cameron, 2008). While diploid interspecific crosses have led to increased biomass yields, recent evidence from family-based selection and yield trials have shown that high-levels of hybrid vigor (heterosis) occurs in interspecific triploids generated from diploid and tetraploid parents (Zsuffa et al., 1984; Serapiglia et al., 2014a; Serapiglia et al., 2014b; Fabio et al., 2016; Carlson and Smart, 2016).

This body of work provides additional evidence that interspecific triploid hybrids of willow (*Salix* spp.) outperform their diploid and tetraploid progenitors, based on the suite of important biomass-related traits collected in the field and greenhouse. Importantly, crosses made between *Salix* individuals from Sections Vimen and Helix showed the greatest levels of heterosis for harvestable biomass, predicted root mass, and stem growth traits. While foliar traits were most susceptible to environmental fluctuation, shape descriptors were more stable between years. Genetic mapping in an association panel and F_2 family of *S. purpurea* was successful in

identifying high-effect QTL linked to biomass, foliar, architecture, composition, physiological, and pathology traits. Sexually dimorphic traits were identified in both *S. purpurea* mapping populations. Along with QTL for sex, foliar traits, and hemicellulose, sexually dimorphic gene expression in *S. purpurea* localized to the large, non-recombining, sex determining region on *Salix* chromosome 15. This was also observed in triploid crosses *S. purpurea* × *S. miyabeana* and *S. viminalis* × *S. miyabeana*.

It was noted that all family 430 individuals were identified as female, whereas both male and female individuals were identified in families 423 and 415, which shared the common tetraploid male parent 01-200-003. Why is it that the sex of the tetraploid parent controls the sex ratios of triploid progeny? For instance, the family 415 cross (*S. purpurea* Z_pW_p × *S. miyabeana* $Z_mZ_mZ_mZ_m$) yields male ($Z_p + Z_mZ_m$) and female ($W_p + Z_mZ_m$) triploid progeny. Likewise, the family 423 cross (*S. viminalis* Z_vW_v × *S. miyabeana* $Z_mZ_mZ_mZ_m$) yields male ($Z_v + Z_mZ_m$) and female ($W_v + Z_mZ_m$) triploid progeny. However, the 430 cross (*S. miyabeana* $Z_mW_mZ_mW_m$ × *S. viminalis* Z_vZ_v) produces all-female ($Z_mW_m + Z_v$ or $W_mW_m + Z_v$) triploid individuals. Why are only females observed in the offspring of tetraploid females? Is it true across all pedigrees? Following quadravalent formation, female tetraploid Z_m and W_m could form Type I tetrads (Z_mZ_m, Z_mZ_m, W_mW_m, W_mW_m) or Type II/III tetrads (Z_mW_m, Z_mW_m, Z_mW_m, Z_mW_m), such that the former would produce progeny segregating with approximately 1:1 sex ratios, but the latter would most likely produce all-females (assuming the male is also tetraploid). Is this reflective of a dosage-effect? Can one W_m chromosome be tolerated in intraspecific tetraploid male offspring (Type II/III), where equal doses of the Z_m and W_m must be preserved for female identity? That still wouldn't account for all-female observations in triploid family 430. Or is it reflective of maternal imprinting, where the sex of the tetraploid parent (when crossed to a diploid) can

change maternal:paternal genome contributions in the triploid endosperm? Diploid female × tetraploid male crosses would result in a balanced tetraploid endosperm (2+2), but a tetraploid female × diploid male crosses would result in a *maternal excess* pentaploid endosperm (4+1). Alternatively, the sex determination systems may differ dramatically between species, which could lead to the loss, misexpression, or incompatibility of a female-suppressing locus or its downstream targets in the triploid hybrid. Nevertheless, a deeper understanding of sex determination among polyploid willows would expand what is known about the evolution of dioecy in the Salicaceae, especially since the genus *Salix* is comprised of a range of ploidy-levels.

For those genes differentially expressed between *S. purpurea* parent shoot tip and internode tissues, expression-level dominance was most prominent, with very few showing additive inheritance. A high proportion of expression-level dominance and additive inheritance was observed among triploid family individuals, and all regulatory patterns were present, with the greatest proportion of allele-specific expression showing *cis*-divergence. Common dominant genes in triploids were enriched for important pathways related to starch and sucrose metabolism. Dosage effects were observed to fit parent genome contributions in triploids, with only a small proportion insensitive to changes. Importantly, the proportion of genes differentially expressed between parents attributable to nonadditive inheritance was most correlated with nonadditive phenotypic expression (heterosis) for biomass yield and stem growth traits among triploid hybrids.

The extensive phenotypes collected in this body of work provides a database of phenotypic information for future trait mapping, genomic selection, and the identification of new crosses for breeding. Gene sets correlated with heterosis for biomass traits in triploid hybrids

help lay the groundwork for increasing biomass yield in different genetic backgrounds. The frequency of expression-level dominance across intraspecific and interspecific diploid and triploid F_1 families, extensive regulatory divergence (especially for higher ploidy levels) offer a unique perspective on the genomic basis of heterosis in outcrossing and highly heterozygous perennial crop plants. In addition, nonadditive gene sets correlated with heterosis for biomass traits will contribute to the growing genomic toolkit for the improvement of willow as a bioenergy crop.

Improvement of shrub willow bioenergy crops in the future will benefit most from exploiting the natural diversity of species and ploidy-levels. Genetic mapping was successful in identifying key bioenergy traits in *S. purpurea*, but the production of large mapping populations for species hybrids and higher ploidy levels would help confirm the results reported here. For instance, there was a general lack of power to detect significant loci associated with biomass-related traits due to the low number of individuals in the *S. purpurea* association panel. Efforts should be made to collect more diverse material, which is likely to be realized via international collaborative support. With the rise in global temperatures, a good area of research would involve the dissection of genotype × environment interactions of key biomass traits, especially in drought-stressed conditions. Environmental factors can greatly impact yields, so if genomic selection becomes an important component of shrub willow breeding in the future, it will be necessary to conduct a greater number of multi-environment trials. Shrub willow is a woody perennial crop, and requires a massive amount of time and effort to phenotype even a single trial. High-throughput phenotyping, like aerial imaging, would drastically improve the feasibility and success of collecting multi-environment data. In the near future, it may be possible to render the

entire field, which could help uncover spatial and genetic variation unobservable to the naked eye.

There is much to learn about the genetic architectures of important biomass and wood chemical composition traits in shrub willow. Significant effort is needed to develop accelerated crop breeding strategies. With the advent of high-throughput, next-generation sequencing technologies and low-cost genotyping protocols, the ability to accelerate breeding and selection, especially with non-model crops, will continue to improve.

7.2 REFERENCES

Argus GW. 1997. Infrageneric classification of *Salix* (Salicaceae) in the New World. *Monographs in Systematic Botany*, 52: 1–121.

Bonosi L, Ghelardini L, Weih M. 2013. Towards making willows potential bio-resources in the South: Northern *Salix* hybrids can cope with warm and dry climate when irrigated. *Biomass and Bioenergy*, 51: 136–144.

Carlson CH, Smart LB. 2016. Electrical capacitance as a predictor of root dry weight in shrub willow (*Salix*; Salicaceae) parents and progeny. *Applications in Plant Science*, 4: e1600031.

Fabio ES, Volk TA, Miller RO, Serapiglia MJ, Gauch HG, Van Rees KCJ, Hangs RD, Amichev BY, Kuzovkina YA, Labrecque M, Johnson GA, Ewy RG, Kling GJ, Smart LB. 2016. Genotype × environment interactions analysis of North American shrub willow yield trials confirms superior performance of triploid hybrids. *GCB Bioenergy*, 9: 445–459.

Serapiglia MJ, Gouker FE, Hart JF, Unda F, Mansfield SD, Stipanovic AJ, Smart LB. 2014a. Ploidy level affects important biomass traits of novel shrub willow (*Salix*) hybrids. *BioEnergy Research*, 8: 259–269.

Serapiglia MJ, Gouker FE, Smart LB. 2014b. Early selection of novel triploid hybrids of shrub willow with improved biomass yield relative to diploids. *BMC Plant Biology*, 14: 74.

Hardig TM, Brunsfeld SJ, Fritz RS Morgan M, Orians S. 2000. Morphological and molecular evidence for hybridization and introgression in a willow (*Salix*) hybrid zone. Mol*ecular Ecology*, 9: 9–24.

Kuzovkina YA, Romero MA, Charles J, Hurst J, Mcivor I, Karp A, Trybush S, Labrecque M, Teodorescu TI, Singh NB, Smart LB, Volk TA. 2008. *Salix*: botany and global horticulture. *Horticultural Reviews*, 34: 447–489.

Lauron-Moreau A, Pitre FE, Argus GW, Labrecque M, Brouillet L. 2015. Phylogenetic relationships of American willows (*Salix* L., Salicaceae). *PLoS ONE*, 10: e0138963.

Lin J, Gibbs JP, Smart LB. 2009. Population genetic structure of native versus naturalized sympatric shrub willows (*Salix*; Salicaceae). *American Journal of Botany*, 96: 771–85.

Smart LB, Cameron KD. 2008. Genetic improvement of willow (*Salix* spp.) as a dedicated bioenergy crop. In: Vermerris W, editor. Genetic Improvement of Bioenergy Crops: Springer New York. pp. 377–396.

Zsuffa L, Mosseler A, Raj Y. 1984. Prospects for interspecific hybridization in willow for biomass production. In: K. Perttu (Ed.) Ecology and Management of Forest Biomass Production Systems. Uppsala, Sweden, Swedish University of Agricultural Sciences. 15: 261–281.